Speaking of landmarks

Bălțărețu, A.A.

Publication date:
2016

Document Version
Publisher's PDF, also known as Version of record

Speaking of landmarks.
How visual information influences reference in spatial domains

Adriana A. Bălţăreţu
Tilburg University

Speaking of landmarks.
How visual information influences reference in spatial domains
A.A. Bălţăreţu

PhD Thesis
Tilburg University, 2016
TiCC PhD series no. 51

This research is funded by The Netherlands Organization for Scientific Research NWO, Promoties in de Geesteswetenschappen, grant number 322-89-008.

Printing was financially supported by Tilburg University.

Printed by: Ridderprint BV, the Netherlands

Speaking of landmarks. How visual information influences reference in spatial domains

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan
Tilburg University op gezag van de rector magnificus,
prof.dr. E.H.L. Aarts,
in het openbaar te verdedigen ten overstaan van
een door het college voor promoties aangewezen
commissie in de aula van de Universiteit
op donderdag 22 december 2016 om 10.00 huur

door

Adriana Alexandra Bălţăreţu,
geboren op 1 februari 1988 te Constanţa, Roemenië

Promotores:	prof. dr. E.J. Krahmer prof. dr. A.A. Maes
Promotiecommissie:	prof. dr. J.A. Bateman dr. A.D.F. Clarke dr. I. Paraboni prof. dr. M.G.J. Swerts dr. M. Theune

Contents

CHAPTER 1

Introduction

Imagine that you are somewhere in an unknown city trying to find a café, but you got lost in the hustle and bustle of the city. Nowadays, we can arm ourselves with an arsenal of navigational tools to prevent such a situation. You can choose between a smart phone, various apps, the Internet, GPS, guidebooks or maps, in order to figure out how to get to your destination. Often these tools distract from what is happening around you; they can occupy all your attention, and you may feel confused about whether you missed that left turn or not. Now, envisage that sometime in the near future, as augmented reality technology is likely to have become interwoven into the fabric of our daily lives, you could put on your headset and you are able to listen to simple instructions that include references to landmark objects and events you see around you: instructions that allow you to connect with the surroundings. Such new technological advances would be able to take into account all visual aspects of the environment, permitting navigation software to generate spoken instructions allowing you to keep your visual attention focused on the real world.

We expect this to become a reality in the near future. For example, augmented reality technology will allow easy capture of the visual environment in real time via small video cameras. This could enable pedestrian navigation systems to generate instructions making use of both stable database information (e.g., streets, reference buildings) and variable visual information captured directly by the camera (e.g., how busy the street is, whether there are moving cars around). Despite the advances in wearable augmented technology, there are still major, technical challenges that the realisation of such a system would pose. Beyond these technical difficulties, it is also not really clear yet what exactly would make a good route direction in such a setting. When should a system refer to an object (e.g., "go left at X") and how should it refer to this landmark X? In some situations, the task of giving good directions might be difficult: should the system adapt the way it refers landmark X? More in general, what types of objects make good landmarks in the first-hand, in-situ experience of the environment? In fact, we still only poorly understand how human speakers produce and understand reference in complex spatial domains. Therefore, one important step towards creating effective instructions is studying how humans make use of space when referring to objects in naturalistic environments. This thesis addresses a specific aspect of the scenario above: the production and comprehension of landmark references in spatial domains, while taking into account the visual context and task-related aspects.

Throughout the four studies reported in this thesis, we discuss various issues related to references in spatial domains. We report on results of psycholinguistic experiments, and we attempt to formulate implications for developing natural language generation algorithms that could automatically produce human-like route directions. The purpose of this introduction is to offer more background information, give an overview of the four studies and explain some methodological aspects. In the next sections, we elaborate on reference in spatial domains and the factors that might affect referential processes.

1.1 The space we live in

Space is a prevalent dimension of everyday life. Human activity obviously takes place in the space we inhabit and through which we navigate. Physical interactions in space trigger representations that are used to support thinking about abstract entities (Lakoff & Johnson, 1980). We make communicative use of space in gestures (Gentner, Özyürek, Gürcanli, & Goldin-Meadow, 2013), in metaphoric thinking ("somebody has fallen into a depression", Lakoff & Johnson, 1980), in actions (lining up the ingredients for a recipe in order of use, Kirsh, 1995) and in reasoning (e.g., using spatial representations when thinking and talking about time, Casasanto & Boroditsky, 2008).

When talking, people frequently refer to space and the objects placed in it. For example, one can ask for "the coffee mug on the table" (identifying a target object, "the mug", in relation to a relatum object, "the table"); give instructions "take a left turn at the blue building on the left" (identifying a landmark, "the blue building on the left", while giving route directions); and make use of visual communication (maps, graphs, etc.) to locate, relate and quantify information (for a review, see Tversky, 2011). Referring expressions such as "the coffee mug on the table" and "the blue building on the left" are frequently present in our conversations, but how human speakers produce and comprehend such expressions that include a spatial component is still largely unexplored.

1.2 Identification in spatial domains

In studies on reference, it is typically assumed that the speaker's purpose is to identify a referent, by means of a particular description, in such a way that the addressee can pick out the intended object (van Deemter, Gatt, van Gompel, & Krahmer, 2012). It often happens that simply naming the entity is not enough (e.g., "the house"), because in the visual context there might be several similar objects (houses) which could fit this simple description. Thus, the speaker needs to choose more properties in order to produce an expression that distinguishes the referent from the rest of the objects (e.g., "the blue house on the left") and doing so sometimes she may add more information than strictly speaking required for identification (e.g., Koolen, Gatt, Goudbeek, & Krahmer, 2011). While referring expressions can come in many flavours (e.g., "the blue house on the left", "it", etc.), this thesis investigates initial definite descriptions which refer to physical objects. The referring expressions analysed typically take the form of a noun denoting the target referent, often coupled with one or more modifiers that can be expressed in different ways (pre-nominal modifiers, e.g., "the blue building", and post-nominal ones, e.g., "the building which is blue").

In this thesis, reference in spatial domains is interpreted on the one hand as the use of relata in locating objects in an environment and, on the other hand, as

the production and comprehension of landmarks (essential elements of spatial mental models, Tversky, 2011). By 'landmarks' we refer to environmental features that may function as points of reference (Allen, 2000). By 'relatum' we refer to objects with respect to which the (position) of a target object is described (Levinson, 1996). Relata are an intrinsic part of 'place descriptions' (Richter & Winter, 2014, p. 92). When a location is being described, it often includes references to objects in the environment that are in some spatial relationship. Consider, for example, spatial descriptions such as "give me the key behind the mug", "turn left at the building next to the restaurant" or "turn left on the street at the taco place". In all cases, in order to localize a target object (the key, the building, the street), the speaker relates it to a second object that is salient in some sense. In this sense, relatum references could be considered similar to landmark references.

Referring expressions have been studied extensively in recent years, most notably in the Referring Expression Generation (REG) and in the psycholinguistic communities (for a review, see van Deemter et al., 2012; Krahmer & van Deemter, 2012). Yet, reference in spatial domains is an underinvestigated topic (for similar arguments, see Dale & Viethen, 2009; Krahmer & van Deemter, 2012; Paraboni & van Deemter, 2014; Paraboni, Galindo, & Iacovelli, 2016). For example, there were various attempts to incorporate relational descriptions into different REG algorithms (Horacek, 1997; Krahmer & Theune, 2002; Kelleher & Kruijff, 2006), often based on the assumption that relational properties ("x is in front of y") are less preferred than non-relational ones ("x is blue"). However, this preference assumption is debatable, for various reasons. Speakers frequently mention the position of a target object, especially when the object is part of a visually complex and naturalistic environment (Viethen & Dale, 2008; Clarke, Elsner, & Rohde, 2013; Kazemzadeh, Ordonez, Matten, & Berg, 2014), although it is not entirely clear what causes speakers to mention the location of the target. This situation might have arisen from the fact that most studies on reference have used relatively artificial tasks (often forbidding the use of locative information) and artificial scenes (grid-like arrangements of objects) (Krahmer & van Deemter, 2012). Therefore, the extent to which visual properties of the objects and of the scene might influence reference in naturalistic environments has been less explored.

Contrastively, in the field of spatial cognition, there have been various studies concerned with principles that define object selection and localization of relata objects (Levelt, 1993; Carlson-Radvansky & Radvansky, 1996; Levinson, 2003; Miller, Carlson, & Hill, 2011; Barclay & Galton, 2008) and landmark objects (for a review, see Richter & Winter, 2014; Tom & Tversky, 2012; Denis, Mores, Gras, Gyselinck, & Daniel, 2014). Yet, these studies mostly focus on how to produce optimal locative information that describes where an object is located, rather than how to use location as one possible attribute among others (such as colour, size, etc.) that uniquely describes an object (for a more detailed argumentation on the differences between 'localization' vs. 'identification' studies, Tenbrink, 2005; Barclay & Galton, 2008; Dos Santos Silva & Paraboni, 2015). Less is known about the distribution of the various

properties (such as colour, size, location) when reference production takes place in naturalistic settings (see also, Clark & Bangerter, 2004).

Determining the content of these descriptions becomes relevant given novel proposals of enriching datasets for automatic route directions with user generated descriptions of landmarks (Richter & Winter, 2014, p. 167). Various studies have examined how humans acquire and use landmarks when new environments are explored (Siegel & White, 1975; Ishikawa & Montello, 2006), yet most studies on route directions do not address how landmarks could be referred to. However, some classifications exist, mostly as a result of earlier qualitative observational work. For example, buildings are often described by proper names of businesses (e.g., "turn left at the Hilton"). Where these name labels are missing, speakers also refer to visual properties of objects for disambiguation purposes (e.g., Elias, Paelke, & Kuhnt, 2005). Landmark references using visual properties of objects (such as "the blue building") pose additional challenges to algorithms relying only on database information, because these databases typically do not store a broad range of perceptual properties of potential landmarks (e.g., Dale, Geldof, & Prost, 2005; Janarthanam et al., 2013; Roth & Frank, 2009). REG insights could potentially be of great help in generating automatic references to landmarks, yet it is an open question to what extent findings from one type of studies, based on references produced in response to a particular task (identification), carry over to other contexts (such as route directions).

In this thesis, we aim to get a better understanding of how people make use of the richness of the visual context and adapt their references to the characteristics of the environment and of the task, with a focus on references to landmarks. This endeavour provides valuable behavioural evidence for developing natural language generation algorithms that could automatically produce human-like route directions.

1.3 Visual properties and task-related aspects

When speakers refer to objects (as landmarks, but also more in general) what they see influences what they say, as has been suggested by many psycholinguistic studies of language production (e.g., Meyer, Sleiderink, & Levelt, 1998; Griffin & Bock, 2000; Hanna & Brennan, 2007). Moreover, it is likely that visual properties influence referential processes (Bock, Irwin, Davidson, & Levelt, 2003; Gleitman, January, Nappa, & Trueswell, 2007; Coco & Keller, 2015). Reference takes place in a complex dynamic environment where (1) not all objects are equally prominent (salient) and (2) where the amount of visual detail affects how easily an object is perceived. In this thesis, we investigate the extent to which visual properties of objects-to-be-described influence relata and landmark references. For example, perceptual salience has been proposed to influence relatum selection (e.g., Barclay & Galton, 2008), and we question to what extent it could affect relatum reference as well. Specifically, when speakers have to refer unambiguously to a target object, we ask what causes speakers to mention

one of the objects as a relatum. Regarding landmarks, there is empirical evidence that attributes such as colour and size influence selection (Raubal & Winter, 2002; Nothegger, Winter, & Raubal, 2004), yet there are also other basic attributes that are processed in the early stages of visual perception (e.g., the direction and velocity of motion, Mital, Smith, Hill, & Henderson, 2011). In particular, motion is a property that has not been thoroughly investigated in the study of navigation communication, and we wonder to what extent it influences reference and the type of objects speakers consider to be relevant when giving live, in situ route directions, where motion is ubiquitous.

Moreover, landmark references may be produced in naturalistic environments which are visually complex and the level of visual clutter (the state in which excess items lead to a degradation of performance at some task, Rosenholtz, Li, & Nakano, 2007) could affect the ease with which speakers uniquely refer to, for example, the street that needs to be taken next. Imagine giving route directions in the busy centre of Berlin, as opposed to giving route directions in a residential suburb. Speakers might have problems giving directions in environments with high levels of visual clutter, and addressees might find it more difficult to find the way. In this thesis, we ask whether and, if so, how speakers tune their references to cope with visual complexity, and to what extent this influences addressee's comprehension and their behaviour.

A speaker does not refer to objects in an empty context, but as part of a larger navigation task, and this could also contribute to the content selection and formulation choices a speaker needs to make. More specifically, we investigate the extent to which the communicative task might influence the production of referring expressions. In this thesis, we focus on two aspects related to the communicative task, namely the purpose of interaction and task complexity.

It has been long argued that the production of referring expressions is sensitive to the communicative context in which they are used (e.g., "make a contribution as informative as required *for the current purposes of the exchange*", Grice, 1975, p.45, our emphasis). However, most existing corpora of referring expressions have been collected (using different instructions) for the purpose of one task only: identification. This raises the question to what extent the insights of these earlier corpus studies generalize to other communicative tasks (like instructing someone). Do speakers tend to use a similar level of specification given different tasks (e.g., identification vs. route directions)? Furthermore, not only the communicative purpose, but also the complexity of the task might influence references in naturalistic settings. Imagine you need to make your addressee take the correct right branch of **K**-shaped intersection as compared to a +-shaped intersection; arguably this would be harder in the former than in the latter case. Task complexity, translated in this navigation scenario as the number and angle of intersection branches, may result in differences regarding referential effort: in complex situations the speaker might be addressee-oriented and give detailed descriptions that suit better the addressee's needs, or not as they try to minimize their own effort.

Using psycholinguistic experiments, we analyse how these factors influence both the production and comprehension of referring expressions in naturalistic environments. Before continuing with an overview of the studies, we address some recurrent methodological aspects.

1.4 Methodology

There are several methodological aspects that are common to the studies presented in this dissertation. Firstly, in each chapter we report on a production study and one or more comprehension or evaluation studies. Analysing both production and comprehension aspects sheds light not only on what objects the speaker chooses and how she refers to them, but also what is effective for the addressee and what the latter prefers. This thesis focuses on production, and takes into account comprehension aspects as a mean to assess the effectiveness of the speaker's contribution.

The production experiments consist of both object identification tasks (Chapter 2 and 3) and object identification while giving route directions (Chapter 3, 4 and 5). Speakers were asked to produce references in such a way that an addressee could identify the objects or give route directions for an addressee that needs to find the correct route. When giving route directions speakers were asked refer to highlighted objects as landmarks (Chapter 3) or had the liberty to decide whether they wanted to add landmark references and to choose the landmark objects (Chapter 4 and 5). Addressees were asked to understand utterances and respond to references (e.g., by clicking on an object, e.g., Chapter 2), choose the correct street (Chapter 4 and 5) or choose the descriptions and route directions they like best (Chapter 2, 3, 4, 5).

Compared to studies that use simple scenes, the level of visual detail in almost all chapters is similar to what human speakers experience on a daily basis. By controlling the type of visual scenes (e.g., the type of intersection, the level of visual clutter), we were able to analyse cause and effect relationships between visual environment and reference, that otherwise would be hard to establish. More specifically, visual properties are hard to measure, manipulate or control when route directions tasks are carried out on the streets (e.g., Lovelace, Hegarty, & Montello, 1999; Denis et al., 2014).

We report on a series of studies that start with a virtual, controlled environment (Chapter 2) and evolve towards naturalistic stimuli which depict almost natural situations (photographs of Google Street View or videos of real intersections, Chapters 3, 4 and 5). It has been argued that the simplicity and artificiality of some earlier referring expressions studies (e.g., TUNA-corpus van Deemter et al., 2012) can be detrimental for the study of visual factors in relation to reference production (Clarke, Elsner, & Rohde, 2013) and could potentially bias the way (psycholinguistically motivated) models of reference production work (Gatt, Krahmer, van Deemter, & van Gompel, 2014; Frank & Goodman, 2012). Compared to these studies, the stimuli used

in these studies dissertation mostly represent real life situations, in which the target objects are an integral part of a (complex) visual scene (rather than being randomly positioned in a grid).

1.5 Focus and Outline

This dissertation reports on four studies related to the production of initial definite references whose content is shaped by information available in the visual context. In the next chapter, Chapter 2, we report the first empirical study, investigating the production of spatial relational descriptions. We question what factors cause speakers to mention one of the objects as (first) relatum, and we analyse the possible influence of the object's spatial position and salience. It is generally assumed that, if an object can grab visual attention, it is salient in some dimension, and is more likely to be selected and mentioned (Beun & Cremers, 1998; Tversky, Lee, & Mainwaring, 1999; Sorrows & Hirtle, 1999; Kelleher, Costello, & van Genabith, 2005; Kelleher & Kruijff, 2006). In a production experiment consisting of several parts, we operationalize the concept of salience in different ways. First, we vary salience systematically by manipulating the conceptual salience of the objects (making one of the relatum candidates animate). Furthermore, we manipulate perceptual salience by adding attention capture cues, first subliminally by priming one relatum candidate with a flash, then explicitly by using salient colours for objects. In a different, acceptability rating experiment, we ask participants to express their preference for specific relata, by ranking descriptions on the basis of how good they think the descriptions fit the scene.

Next, in Chapter 3, we questioned to what extent findings from one field (identification studies) generalize to a different context (route directions). Typically, in identification studies, the purpose for which the speaker produces a referring expression is to identify an object for an addressee (Krahmer & van Deemter, 2012); while in route directions, objects are being referred to in the light of a more complex task, such as finding the correct street. The purpose of the interaction introduces a specific perspective of the situation (such as describing or instructing), which could influence the level of informativeness of a contribution (e.g., Clark, 1996's work on dialogue). In the third chapter, we contrast two tasks with different purposes: identification and instruction giving. In one production experiment, speakers referred to a target building nearby or further away, so that their addressee would distinguish it between other buildings (identification) or give route directions and use the same building as a landmark (instructions). Next in an evaluation experiment, participants were presented with both references produced in the identification condition and in the route directions one, and had to choose the best matching reference, while thinking that they are evaluating descriptions of objects or descriptions of objects extracted from route directions.

In Chapter 4, we zoom in on the question whether visual properties of the scene

affect reference production and comprehension of route directions. We focus on two aspects of the visual surroundings, namely a perceptual factor, visual clutter, and a task related factor, the intersection structure. Visual clutter has been found to affect, for example, object recognition performance (Bravo & Farid, 2006), scene segmentation (Bravo & Farid, 2004), and visual search (Henderson, Chanceaux, & Smith, 2009). Recently, clutter has been shown to affect not only scene perception, but also language and reference production (Coco & Keller, 2009; Koolen, Krahmer, & Swerts, 2013; Clarke, Elsner, & Rohde, 2013). In a visually noisy environment, speakers might have problems giving route directions (which we address in Experiment 1) and addressees might find it more difficult to find the way (Experiment 2). Moreover, the inherent complexity of the task could influence how well speakers cope with increased difficulty (Experiment 1) and if their strategies are beneficial for the addresses and also preferred by the latter (Experiment 2 and 3).

In Chapter 5, we continue exploring the relation between movement, a factor contributing to the perceptual salience of objects and reference production in the context of route directions. The motivation for this study comes from one of the findings presented in Chapter 4, namely that speakers regularly choose 'non-typical' landmark objects, such as parked or moving cars and pedestrians. Based on earlier literature on landmarks, such choices might seem surprising, yet moving objects might be natural points of reference for people in live situations, since movement is one of the features that contribute to perceptual salience. In this chapter, we therefore investigate if and when speakers refer to moving entities in route directions and how listeners evaluate such instructions. We asked speakers to watch short videos of different crossroads with and without moving landmarks and give directions to listeners, who in turn had to choose a street on which to continue (Experiment 1) or choose the instruction they most preferred among three directions (Experiment 2).

The last chapter of this thesis contains a general discussion and final conclusions. Though these chapters taken together form a larger story, they can be also read on their own. Three chapters are based on articles that have been published in scientific journals, and one is currently under review. The chapters are self contained texts, and all have their own abstract, introduction, and discussion. Due to this self-contained nature of the chapters, a small amount of redundancy was unavoidable. As a result of different requests from different reviewers and journals editors, the studies reported make use of slightly different methods and techniques of data analysis, phrasing and presentation of results. The author of this thesis was the main researcher in all the studies reported.

CHAPTER 2

Talking about Relations: Factors influencing the production of relational descriptions

Abstract In a production experiment (Experiment 1) and an acceptability rating one (Experiment 2), we assessed two factors, spatial position and salience, which may influence the production of relational descriptions (such as "the ball between the man and the drawer"). In Experiment 1, speakers were asked to refer unambiguously to a target object (a ball). In Experiment 1a, we addressed the role of spatial position, more specifically if speakers mention the entity positioned leftmost in the scene as (first) relatum. The results showed a small preference to start with the left entity, which leaves room for other factors that could influence spatial reference. Thus, in the following studies, we varied salience systematically, by making one of the relatum candidates animate (Experiment 1b), and by adding attention capture cues, first subliminally by priming one relatum candidate with a flash (Experiment 1c), then explicitly by using salient colors for objects (Experiment 1d). Results indicate that spatial position played a dominant role. Entities on the left were mentioned more often as (first) relatum than those on the right (Experiment 1a, 1b, 1c, 1d). Animacy affected reference production in one out of three studies (in Experiment 1d). When salience was manipulated by priming visual attention or by using salient colors, there were no significant effects (Experiment 1c, 1d). In the acceptability rating study (Experiment 2), participants expressed their preference for specific relata, by ranking descriptions on the basis of how good they thought the descriptions fitted the scene. Results show that participants preferred most the description that had an animate entity as the first mentioned relatum. The relevance of these results for models of reference production is discussed.

This chapter is based on:
Baltaretu, A., Krahmer, E., Maes, A., & van Wijk, C. (2016). Talking about relations: Factors influencing the production of relational descriptions. *Frontiers in Psychology, 7* (103). doi: 10.3389/fpsyg.2016.00103

2.1 Introduction

Human speakers have a rich repertoire for referring to objects in visual scenes. For example, if you want to buy a ball from the toy store, the shop assistant could help you find it among other balls by referring to intrinsic attributes (e.g., color, *the red ball*) or extrinsic ones (e.g., location, *the ball between the doll and the train*). An object's location can be described in relation to one's body and to other objects or to environmental features (Levinson, 1996). In this chapter, we focus on referential choices when describing external relations (Levinson, 2003; Tenbrink, 2011) where an object is the target, while other object(s) serve as the relatum. The target is sometimes referred to as the locatum, figure or located object, whereas the relatum is also known as ground, reference location or landmark. In the previous example, the ball represents the target and it is described in relation to two relata objects, the doll and the train.

Compared to intrinsic attributes (such as colour), there are few studies in the referring expressions generation field analysing how extrinsic attributes (such as location) are used in order to refer unambiguously to a target object (for a review, see Krahmer & van Deemter, 2012). When talking about location, speakers describe where the target object is positioned in space. Far from being a trivial feature, space is a pervasive dimension in language and cognition. For example, we map time onto space (e.g., Boroditsky, 2000), make use of space in gestures (e.g., Gentner et al., 2013), in discourse (e.g., Lakoff & Johnson, 1980), and in actions (e.g., Kirsh, 1995). Crucially, humans employ location in a meaningful way in different forms of descriptions and visualizations. It is natural to refer to an object's location in a variety of situations, thus anchoring the conversation topic in the spatio-temporal context (Levelt, 1993, p. 51). Such situations are, among other things, route direction production, interaction with conversational agents, visual communication (e.g., maps and graphs) within various disciplines (e.g., architecture, geosciences, engineering, etc.), (for a review, see Tversky, 2011).

Pervasive use of spatial relations in real life communication makes it necessary to develop referring expression generation algorithms that can handle such reference. These algorithms (e.g., the Incremental Algorithm, Dale & Reiter, 1995; the Graph-Based Algorithm, Krahmer, Van Erk, & Verleg, 2003) have a key role in natural language generation, enabling machines to make informed choices and to refer to objects in a more human-like manner (Dos Santos Silva & Paraboni, 2015; van Deemter et al., 2012; Gatt et al., 2014). Though we know little of the situations when relational descriptions are spontaneously produced and preferred over intrinsic attributes, there are communicative contexts in which relations are an efficient and relevant strategy (like in route directions or in scenes with many (similar) objects). Recent studies have shown that speakers often produce relational descriptions in order to single target objects out of other objects in a visual scene (Kazemzadeh et al., 2014; Clarke, Elsner,

& Rohde, 2013). When both intrinsic and extrinsic attributes are available, people tend to mention location even when this attribute is not necessary for producing a unique object description (Viethen & Dale, 2008). Listeners seem to benefit from this type of reference as well (Paraboni & van Deemter, 2014; Arts, Maes, Noordman, & Jansen, 2011). Currently, spatial relations represent a major challenge for referring expressions generation algorithms, as we know little about the situations in which speakers employ them in the context of identification. To further develop these algorithms, more input from studies on human reference is needed.

In this chapter, we focus on human reference production in spatial relational descriptions. In visual scenes, several entities can be in the proximity of the target and each one of them could be a potential relatum. In our previous example, the shop assistant could either refer to the target as, for example, *the ball in front of the doll* (using a single relatum) or *the ball between the doll and the train* (using two relata). In the first description, which we call *the single-relatum formulation*, the question is what causes speakers to mention one of the objects. In the second strategy, *the two-relata formulation*, we question what causes speakers to mention one of the objects as first relatum. In the two-relata formulation, we consider important the order in which entities are mentioned. Word order choices have been previously suggested to reflect speaker's referential preferences (Goudbeek & Krahmer, 2012) and the ease with which these entities are processed (Bresnan, Cueni, Nikitina, & Baayen, 2007; Onishi, Murphy, & Bock, 2008; Jaeger & Tily, 2011).

While the study of spatial relations in the field of referring expression generation is a topic largely unexplored, in the field of spatial cognition there have been numerous studies concerned with principles that govern relatum object selection (e.g., Miller et al., 2011; Barclay & Galton, 2008, 2013), the choice of adequate spatial prepositions based on geometric and functional characteristics of the objects (e.g., Carlson-Radvansky, Covey, & Lattanzi, 1999; Coventry & Garrod, 2004) and the influence of frames of reference on relatum selection (e.g., Levinson, 2003; Tenbrink, 2007; Carlson-Radvansky & Radvansky, 1996; H. Taylor & Rapp, 2004). Various factors might affect the selection of a relatum object. Compared to target objects, relata are described as larger, closer to the target, geometrically more complex (Barclay & Galton, 2013) as well as more familiar, expected, more immediately perceivable (Talmy, 2003).

In this chapter, we seek to investigate speakers' referential choices, aiming thereby to provide further insight for REG algorithms. Most studies mentioned above focus on the problem of localization, as opposed to identification (Tenbrink, 2005; Dos Santos Silva & Paraboni, 2015). In localization tasks speakers are restricted to refer to already agreed upon objects (e.g., the target and relatum are given and a priori labelled as, for example "cup"), based solely on their spatial locations. On the other hand, freely producing a referring expression (like "the cup between the plate and the kettle") is a matter of choosing target attributes (including its spatial position), to help the addressee identify a target object out of several candidates. Comparisons be-

tween identification and localization tasks have been previously addressed (Tenbrink, 2005; Moratz & Tenbrink, 2006; Vorwerg & Tenbrink, 2007). In general, descriptions seem to be more detailed when the target needs to be localized, rather than identified. Factors to influence reference production (e.g., spatial biases, conceptual and visual salience) have been addressed to a lesser extent.

It is generally assumed that if an object is salient, it can grab visual attention, and thus is likely to be selected and mentioned as relatum (Tversky et al., 1999; Beun & Cremers, 1998). A number of visual factors have been identified as important cues for salience, such as size, color, orientation, foregrounding, animacy (for a review, see Wolfe, 1994; Coco & Keller, 2015; Kelleher et al., 2005; Parkhurst, Law, & Niebur, 2002), but little is known about how these and other cues influence reference production. The goal of the current research is to examine two factors previously shown to influence language production and comprehension in general, yet understudied in reference production: spatial position and salience.

2.1.1 Spatial position: a left-to-right preference?

Referring to a relatum may be influenced by a factor present in any visual scene: the position of the object in the scene. Different types of evidence suggest there might be a bias to choose objects placed in specific locations. Speakers choose and mention spatially aligned and proximate objects as relata (e.g., Craton, Elicker, Plumert, & Pick, 1990; Hund & Plumert, 2007; Miller et al., 2011; Viethen & Dale, 2010). Yet, when several objects are in the vicinity of the target, all similarly aligned, would spatial features continue to influence reference production? We assume that it does, and objects on the left of the target would be mentioned more often as relatum than objects on the right. This prediction is based on findings from various disciplines.

The speaker's attention might be guided by different factors towards specific regions of the scenes. One line of research suggests that oculomotor biases (the amplitude and direction of saccades - movements of the eye between fixation points) are an important predictor for the location where speakers initially direct their attention (e.g., Tatler & Vincent, 2009; Kollmorgen, Nortmann, Schröder, & König, 2010). One well known, image independent bias is the tendency to look at the centre of visual stimuli during image exploration (for a review, see Clarke & Tatler, 2014). Besides this bias, there is also evidence for a horizontal spatial bias (sometimes referred to as "pseudoneglect"). People initially execute more often leftward than rightward saccades, irrespective of the content of the image, across different tasks (free viewing, memorization, scene search, Ossandón, Onat, & König, 2014; Foulsham, Gray, Nasiopoulos, & Kingstone, 2013). This asymmetry seems to affect memory, with left positioned objects being better remembered than right positioned ones (Dickinson & Intraub, 2009).

Converging evidence comes from cross-cultural psychology research where the left-to-right bias is considered to be a result of the scanning routines employed during

reading and writing. The directionality of the language system has an impact on visual attention, memory, and spatial organization (T. T. Chan & Bergen, 2005). For instance, when participants with a left-to-right language system (in this case: French) were asked to mark the middle of a straight line, they usually misplaced the mark to the left of the objective middle, while participants with a right-to-left language system (Hebrew) misplaced the mark to the right (Chokron & Imbert, 1993). Such a bias is shown from a young age in graphical representations of spatial and temporal relations (Tversky, Kugelmass, & Winter, 1991). This implies that, at least in western cultures, people 'read' visual scenes from left to right and that the left-to-right bias might be a habit acquired by systematically using a language system.

The directionality of the writing system seems to affect cognitive linguistic processes. In picture description tasks, speakers of left-to-right languages tend to scan, describe and remember items from left to right (Taylor & Tversky, 1992; Meyer et al., 1998). Speakers of different writing systems show different patterns of sentence production. For example, in a sentence-picture matching task, speakers of a language with a left-to-right (in this case: Italian) system tended to choose visual scenes with the agent placed on the left of the patient, those of a language with a right-to-left system (Arabic) preferred scenes with the agent placed on the right of the patient (T. T. Chan & Bergen, 2005; Maass & Russo, 2003). Not only the writing system, but also the dominant frame of reference of the language, might affect the order in which speakers refer to entities in visual scene. For example, when using a relative frame of reference, to perceive that something is 'on the left', the speaker would project his viewpoint onto the scene (Levinson, 2003). Bilingual speakers of Spanish (a language with a relative frame of reference) and Yucatec (a language with no dominant frame of reference), show a bias to start with the left object in the scene when using Spanish, but not when doing this task in Yucatec (Butler, Tilbe, Jaeger, & Bohnemeyer, 2014).

The left-to-right bias was also observed in clinical populations. Participants suffering from agrammatism, an aphasic syndrome, presented a similar left-to-right bias both in language production (describing visual scenes) and comprehension (matching sentences with pictures) (Chatterjee, 2001). In addition, studies in the psychology of art suggest that reading habits influence visual preferences: participants preferred pictures possessing the same directionality as their reading system (Chokron & De Agostini, 2000).

Given the evidence for a left-to-right bias, there might be a tendency for speakers to mention relata based on their position in the scene. For example, in Figure 1, speakers could refer to the target as in a) *the ball in front of the bookshelf*, b) *the ball in front of the clock* or c) *the ball between the bookshelf and the clock*. These three descriptions were considered valid for identification and classified in two formulation preferences: the single-relatum formulation (descriptions a and b) and the two-relata formulation (description c). When only one object was mentioned, we considered it to reflect the speakers' preference for a relatum candidate. In case both entities

were mentioned, we took into account the order of mentioning. If a left-to-right bias plays a role in reference production, we expect entities left of the target to be mentioned more often as relatum (as in *a*) or mentioned more often as the first relatum (as in *c*). However, a spatial bias, might not be the sole factor that influences relatum reference. In the following section, we review evidence for other factors that potentially contribute to the salience of relatum candidates.

2.1.2 Salience

Salience is generally considered an important factor for reference production. The objects' salience captures visual attention and entities in focus of attention during utterance planning have higher chances of being mentioned (Gleitman et al., 2007; Beun & Cremers, 1998). In the present chapter, salience (the property of being noticeable or important) is operationalized in two ways.

We distinguish between conceptual and visual salience. By conceptual salience, we refer to the ease of activation of mental representations caused by knowledge-based conceptual information (or 'accessibility' in Ariel, 1990; Bock & Warren, 1985). There are several properties of the referent that contribute to its conceptual salience (e.g., linguistic properties, such as the syntactic position a referent occupies; context, such as the preceding discourse; intrinsic properties, such as animacy, etc.). In this chapter, we focus on animacy: whether an entity is conceptualized as living or not (Vogels, Krahmer, & Maes, 2013; Coco & Keller, 2015). In contrast, by visual salience we touch on two different aspects: perceptual salience and visual priming. By perceptual salience, we refer to bottom-up, stimulus-driven signals that attract visual attention to areas of the scene that are sufficiently different from the surroundings (Itti & Koch, 2001). For example, a perceptually salient object is an object that has a unique color compared to the rest of the scene. Moreover, entities can become salient when visual attention is guided towards them, for example by using attention priming techniques (Gleitman et al., 2007). Below we discuss these types of salience in more detail.

Conceptual salience

Animacy is a basic conceptual feature of objects and there are reasons to believe that it may affect the production of relational descriptions. First, animacy has been shown to influence the allocation of visual attention. Humans prioritize the visual processing of animate objects over inanimate ones (Fletcher-Watson, Findlay, Leekam, & Benson, 2008; Kirchner & Thorpe, 2006; New, Cosmides, & Tooby, 2007). Both visual representations of the face and the human body have the ability to capture the focus of attention, even when attention is occupied by another task (Downing, Bray, Rogers, & Childs, 2004). Compared to inanimate objects, animate entities are more likely to be fixated and named (Clarke, Coco, & Keller, 2013; for a review, see Henderson & Ferreira, 2013).

Second, animacy is known to play a key role in reference production (McDonald, Bock, & Kelly, 1993; Clark & Begun, 1971). Animate entities are conceptually highly accessible, thus, retrieved and processed more easily than inanimate entities (Prat-Sala & Branigan, 2000). This can influence word ordering, as there is a strong tendency for the animate entities to occupy more prominent syntactic positions (e.g., in the beginning of a structure) and grammatical functions (e.g., subject role) (e.g., Branigan, Pickering, & Tanaka, 2008; Prat-Sala & Branigan, 2000; McDonald et al., 1993; Bock, Loebell, & Morey, 1992). Additionally, compared to inanimate referents, animates are mentioned more frequently and are more likely to be pronominalized (e.g., Fukumura & van Gompel, 2011).

Given that utterance planning is influenced by conceptual factors and that animacy has a privileged role in language production, we could expect animate entities to be mentioned as relatum (or as first relatum) more often than inanimate ones due to their conceptual salience, irrespective of their position with respect to the target. In general, there is little evidence that animacy could influence relatum choice. The few studies that looked at this, directly or indirectly, do not present a consistent picture. Under specific circumstances, (de Vega, Rodrigo, Ato, Dehn, & Barquero, 2002) report that relata can be animate, but only when included in a construction using the preposition *behind* [the animate entity]. Congruent evidence was found in a large English corpus of referring expressions elicited with complex naturalistic scenes. Speakers were shown an image with an outlined object and provided with a text box in which to write a referring expression. When speakers decided to produce spatial relational descriptions, the most frequent relata objects were people and some entities positioned in the background, such as trees and walls (Kazemzadeh et al., 2014). T. Taylor, Gagné, and Eagleson (2000), however, argue that animate entities should be disfavored as relata due to their mobility.

Visual salience

Reference production was shown to be sensitive to both visual priming (e.g., a short flash at the target location, Gleitman et al., 2007) and perceptual salience cues, such as uniquely colored objects (Pechmann, 1989; Belke & Meyer, 2002).

Priming participants' initial gaze to a specific area of a scene has been claimed to influence grammatical role assignment and word order (Gleitman et al., 2007). When visual attention is guided towards it, an object is more likely to be mentioned in the beginning of a description or relation (in a prominent grammatical role, such as subject, or in a prominent position in the utterance). As far as we know, no studies looked into effects of attention manipulation on spatial relational descriptions. Reference production can be influenced by very basic, implicit attention-grabbing cues. Gleitman et al. (2007) report that presenting a flash shortly before displaying a scene, systematically redirected the gaze of the participants to the location of a specific object (occurring at the location of the flash), which later received a privileged position

in the sentence structure. The short duration of the flash ensured that participants remained unaware of the manipulation, while their gaze was attracted to the cued location in an implicit manner.

A similar approach has been used for the study of spatial relational descriptions (*X is left of Y*). Forrest (1996) drew speakers' attention to the location of an object, prior to the scene presentation. Unlike Gleitman et al. (2007), she used an explicit visual cue, a flash that lasted long enough to be noticed by the participants. This explicit visual cue influenced speakers' description as well: the object which appeared in the primed location generally received a more prominent place in the beginning of the sentence.

Apart from priming, properties of the stimulus may play a crucial role in guiding the eyes. Perceptual salience is a factor known to influence visual attention (for review, see Tatler, Hayhoe, Land, & Ballard, 2011) and reference production (Coco & Keller, 2015; Clarke, Coco, & Keller, 2013; Myachykov, Thompson, Scheepers, & Garrod, 2011). Perceptual salience is a characteristic of parts of a scene (objects or regions), that appear to stand out relative to their neighbouring parts and there are several models to account for this phenomenon (for a review, see Borji & Itti, 2013). Most models use image features, such as color, contrast, orientation and motion and make center-surround operations to compare the statistics of image features at a given location to the statistics in the surrounding area (Borji & Itti, 2013).

Among these features, colour has been shown to capture visual attention (Folk, Remington, & Wright, 1994; Parkhurst et al., 2002), irrespective of the observers' task (Theeuwes, 1994). In general, colour enhances object recognition (for a review, see Tanaka, Weiskopf, & Williams, 2001) and uniquely coloured items are detected faster than other objects in the scene, regardless of the amount of distractors (Treisman & Gelade, 1980; D'Zmura, 1991).

In general, scholars suggest that explicit perceptual features (such as colour, size, shape) may contribute to relatum selection (e.g., Barclay & Galton, 2008), yet there are almost no experimental studies which try to disentangle the effects of these features. Regarding the influence of colour on relatum selection and reference, prior results are equivocal (Miller et al., 2011; Viethen, Dale, & Guhe, 2011). Yet, in reference production studies, colour is probably the attribute mentioned most frequently. In reference tasks, colour is considered to have a high pragmatic value (Davies & Katsos, 2009; Belke & Meyer, 2002). Speakers mention it even when this information is not needed for identification (Koolen et al., 2011; Westerbeek, Koolen, & Maes, 2015). In complex scenes, reference to both target and relatum objects is affected by perceptual salience (a composite measure of colour and other low level visual features), visual complexity (clutter), size and proximity (Clarke, Elsner, & Rohde, 2013). Clarke, Elsner, and Rohde (2013) note that relatum objects were chosen based on their size and saliency; while references to less salient target objects included a higher number of relata.

Moreover, the order in which objects are mentioned in a relational description

may be sensitive to perceptual salience as well. In visual domains, speakers can mention target and relatum objects in different orders. Elsner, Rohde, and Clarke (2014) report that speakers employed complex word orders such as starting with a) the target, b) the relatum or by giving information about the target in multiple phrases intertwined with relatum references. For example, if the target was a person (target in **bold**, relatum in *italics*), speakers could say a) **man** closest to *the rear tyre of the van*, b) near *the hut that is burning*, there is **a man holding a lit torch in one hand, and a sword in the other** or c) there is **a person standing** in *the water* **wearing a blue shirt and yellow hat** (Elsner et al., 2014, p. 522). These relations were more likely to start with the perceptually salient object.

Given these findings, we could expect objects to be mentioned as (first) relatum if they are placed in a cued location or if they are perceptually salient.

2.1.3 The current experimental studies

Spatial position (left-to-right bias), conceptual salience (animacy), and visual salience (attention capture cues or scene based perceptual cues) all influence what is being looked at (Kollmorgen et al., 2010) and possibly mentioned (Coco & Keller, 2015). We study if and to what extent these factors influence referential choices in spatial relational descriptions.

This chapter presents two experiments consisting of several parts that test the influence of these factors on relatum reference in an identification task. In Experiment 1a, we started by determining if there was a spatial bias when mentioning a relatum. We start with a basic language elicitation task that did not include any experimental factors. Its purpose was to check for a left bias in reference production. In this language elicitation task, we manipulated the position of two inanimate relatum candidates. Entities placed on the left of the target were expected to be mentioned as (first) relatum more often than those placed on the right. We took spatial position as a baseline and continued investigating the effect of salience on referential choices. Conceptual salience was manipulated by adding one animate entity in each scene (Experiment 1b). Animate entities were expected to be preferred as relatum. Visual salience was manipulated by priming attention towards a relatum candidate with a short flash (Experiment 1c) or explicitly with a unique colour (Experiment 1d). Salient entities were expected to be preferred as relatum. Additionally, the listeners' preference for relata was tested, by asking participants to rank relational descriptions starting with the one that, according to them, "best fits" the scene (Experiment 2). Descriptions that have an animate entity as (first) relatum were expected to be ranked higher.

We explored these predictions across a production experiment (four parts) and in an acceptability rating experiment, and in doing so some factors may be included in several parts of these experiments (for example, the effect of spatial position is analysed in Experiment 1 and 2, animacy in Experiment 1b–1d and in Experiment 2,

visual salience in Experiment 1c–1d). Whether speakers mentioned the left entity as (the first) relatum was tested by comparing the chance of naming the left item with random chance (0.50) using an one-sample t–test and possible interactions between the experimental factors were evaluated using analysis of variance (ANOVA) tests [1].

Finally, the current studies were carried out in accordance with the recommendations of APA guidelines for conducting experiments, the Netherlands Code of Conduct for Scientific Practice and the Code for Use of Personal Data in Scientific Research (KNAW). The studies were approved by the ethics committee at Tilburg University and all participants gave written consent to the use of their data.

2.2 Experiment 1 - Reference Production

2.2.1 Experiment 1a - Position

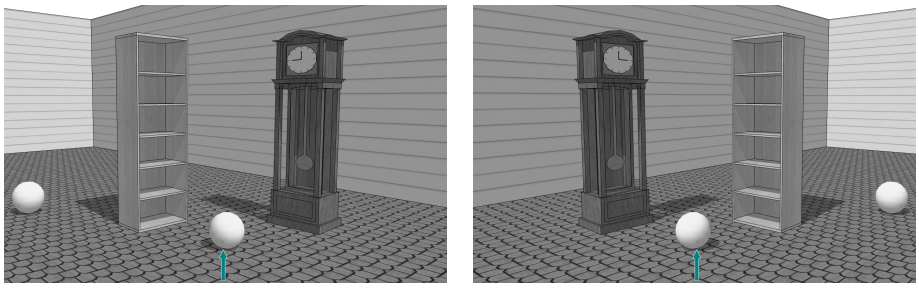
Participants

Thirty native Dutch undergraduates from Tilburg University participated in this study for partial course credits. Data from four speakers were discarded on the basis of task misunderstanding. The final sample consisted of 26 participants (11 female, mean age 20.19).

Materials

The stimuli consisted of 48 greyscale scenes (12 experimental stimuli). The experimental stimuli scenes included a target item marked with an arrow (a ball), a distractor object (a ball identical to the target) in order to prevent an easy identification strategy using type only, and two relatum candidates (both inanimates). These items were eight everyday objects (such as wardrobes), easily identifiable, with a clear front/back axis and of roughly equal size, randomly coupled in pairs (see Figure 1). Filler stimuli were used to have a larger visual diversity (they included both inanimate and animate objects) and to allow participants to use a wider range of identification strategies (type, location and size). All the objects (8 animate and 8 inanimate) were pretested with a group of ten participants, who were presented with pictures similar to the ones used in this study. They had to name the inanimate objects, as well as the gender and profession of animate objects. An inanimate object was included in the experimental stimuli if (1) it was referred to with the same noun in a minimum of 50 percent of the cases, and (2) if the other nouns used to refer to it, were compound nouns such as in "kast"–"ladenkast" (drawer). An animate object was chosen if (1) the character's gender was recognized in all cases and (2) if the character's profession was recognized

[1] The Huynh-Feldt epsilon value was pretty close to 1 in all the analyses, indicating that there was no need for adjustments of the degrees of freedom.



(a) (b)

Figure 2.1: Experimental stimulus with inanimated object (bookshelf) on the left (a) and the right (b) of the target

in 80 percent of the cases. The scenes were created using Google SketchUp 8 (3D Warehouse library).

Procedure

Participants were instructed to verbally refer to an object marked with an arrow in such a way that the next participant (a fictitious listener) could draw the arrows on a new set of identical pictures (language: Dutch). The goal of this instruction was to avoid participants to produce ambiguous references (for a similar procedure see Koolen et al., 2011; Clarke, Elsner, & Rohde, 2013). Participants saw each entity in three different pictures, paired every time with a different object. The materials were divided across two presentation lists, so that each participant would see each object combination only once. The position of each object and the position of the distractor ball were individually counterbalanced (half of the times they appeared on the left of the scene and half of the times on the right of the scene). Descriptions such as *the ball in front of me* or *the ball on the left* were discouraged, by telling the speaker that the listener would receive the same image, but that it might be in a mirror version. The picture remained on the screen until the participant produced a description and pressed a button to continue. Each experimental trial was followed by 3 filler trials to prevent a carry-over effect. The study started with 3 practice trials followed by 48 experimental trials and lasted approximately 10 minutes.

Results and Discussion

We collected 312 descriptions (26 participants * 12 experimental stimuli). Participants were found to use one of two possible formulations: either mentioning a single relatum (e.g., *the ball in front of the bookshelf*) or both (e.g., *the ball in between the bookshelf and the clock*). In all the studies of Experiment 1, the participants were grouped

based on their preference for the single-relatum or the two-relata formulation strategy. Some participants systematically used a single formulation strategy, while others used both. The grouping threshold was set by inspecting the distribution of the two-relata formulation in Experiment 1. The distribution appeared to be bimodal: one group had a score of maximum 100 percent (down to 80); the other group had a score of maximum 40 percent (down to 0). Every participant with a score of 80 or more was considered to opt for a two-relata formulation and all the other for a single-relatum formulation.

In Experiment 1a participants were found to use a single-relatum formulation ($N = 1$ participant, not analysed further due to small sample size) or a two-relata formulation ($N = 25$ participants). Whether speakers mentioned the left entity as the first relatum was tested by comparing the chance of naming the left item with random chance (0.50) using an one-sample t–test. Speakers mentioned the left entity as first relatum 59 percent of the time (95% CI [0.525; 0.659], $SD = 0.16$). This result was statistically significant ($t(24) = 2.857$, $p = .009$; $d = 0.57$).

The results showed a left bias in reference production, however there was only a small preference in starting with the left entity. This leaves room for other factors that could influence reference. Thus, in Experiment 1b, 1c, and 1d, we added three experimental factors that contribute to the entity's salience, making the entities 'stand out' in the scene.

2.2.2 Experiment 1b - Conceptual salience: Animacy

Participants

Fifty three native Dutch undergraduates from Tilburg University participated in this study as speakers for partial course credits. Due to technical problems, speech data of four participants were not analysed; the final sample included 49 participants (11 males, mean age 21.2 years).

Materials

The stimuli consisted of 96 greyscale scenes (24 experimental stimuli). For these scenes, we used the same animate and inanimate objects described in Experiment 1a. The experimental stimuli consisted of a target and a distractor ball and two relatum candidates, one animate and one inanimate object of roughly equal size (see Figure 2). From 64 possible animate–inanimate combinations, 24 couples were randomly chosen. Filler stimuli were similar to the ones used in Experiment 1a.

Procedure

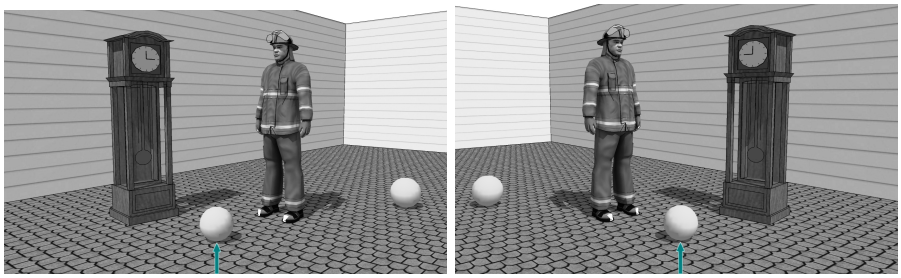
As in Experiment 1a.

Results and Discussion

Speakers produced 1176 descriptions (49 participants * 24 experimental stimuli). Participants were found to use one of two possible formulations: either mentioning a single relatum ($N = 12$) or both relata ($N = 37$). Whether speakers mentioned the left entity as the first relatum was tested by comparing the chance of naming the left item with random chance (0.50) using an one-sample t–test. The chance of mentioning the left entity as first relatum was 59 percent (two-sided 95% CI [0.55, 0.64], $SD = 0.17$, $t(47) = 3.91$, $p < .001$, $d = 0.75$).

Whether animacy overruled the left bias was tested with an ANOVA test, having Position of the Animate in the scene (2 levels: animate left, animate right) as a within subjects factor, and Participant Formulation Preference (2 levels: single-relatum, two-relata) as a between subjects factor. The ANOVA test revealed no statistically significant effect of Position of the Animate ($F < 1$) or of Participant Formulation Preference ($F < 1$) and no interaction between these factors ($F < 1$).

These results suggest that animacy did not influence descriptions. The responses were not affected by word frequency: 90 percent of the participants referred to the animate entity using highly frequent words such as *de vrouw* / *de man* (the woman / the man). However, the position of the entity was found to affect reference to a greater extent, with left entities being more likely to be mentioned as (first) relatum than right ones. In Experiment 1c, we test the strength of this preference by manipulating the objects' visual salience.



(a) (b)

Figure 2.2: Experimental stimulus with animated object (firefighter) on the right (a) and the left (b) of the target object

2.2.3 Experiment 1c - Perceptual salience: Flash

Participants

Thirty nine native Dutch undergraduates from Tilburg University participated in this study for partial course credits. Data from 27 participants (18 women, mean age 20.3 years) were used, the rest being discarded on the basis of having noticed the cue (1 participant), task misunderstanding (2 participants) or not using a relatum at all as in *the ball in the center* (9 participants).

Materials

Stimuli from Experiment 1b were used, slightly cropped so that the target object was placed exactly in the middle of the scene. The attention capture manipulation consisted of a black square, with an area of 0.5×0.5 degrees of visual angle, set against a white background (Gleitman et al., 2007).

Procedure

The procedure was identical to the one presented in Experiment 1a. In addition, an implicit visual attention cue was added. Participants sat approximately 60 cm from the monitor, set to 1680×1050 pixels, 60 Hz refresh rate. Before each trial, participants were first presented with a fixation cross on a white background (500ms). The fixation cross was followed by the attention capture manipulation, which was presented for 65ms, followed immediately by a stimulus scene. The position on screen of the attention-capture cue varied (in half of the trials the cue was positioned left and in half right).

Results and Discussion

Participants used one of the two formulations (single-relatum $N = 6$, two-relata $N = 21$). Whether spatial position influenced reference production was tested by comparing the chance of mentioning the left entity as first relatum with random chance, using one–sample t–test. The chance of mentioning the left entity as first relatum was 67 percent (two-sided 95% CI [0.59, 0.75], $SD = 0.19$, $t(26) = 4.61$, $p < .001$, $d = 0.67$).

Whether animacy or attention priming overruled the left bias was analysed with an ANOVA test, having the Position of the Animate (2 levels: animate left, animate right) and the Position of the Flash (2 levels: flash left, flash right) as within subjects factors, and Participant Formulation Preference (2 levels: single-relatum, two-relata) as a between subjects factor. The ANOVA test revealed no statistically significant main effects of the Position of the Animate ($F < 1$) or of the Position of the Flash ($F < 1$).

There was a main effect of Participant Formulation Preference ($F(1, 25) = 6.66$, $p = .016$, $\eta_p^2 = .21$). In the two-relata formulation, participants mentioned more often the left entity as (first) relatum ($M = .72$), than in the single-relatum formulation ($M = .51$). There were no significant interactions between these factors ($F < 1$).

Experiment 1c confirmed the speaker's preference to mention left entities first. There were no effects of the Position of the Animate or of the Position of the Flash. In Experiment 1d, we continue testing the strength of the left bias by making one of the entities perceptually salient.

2.2.4 Experiment 1d - Perceptual salience: Color

Participants

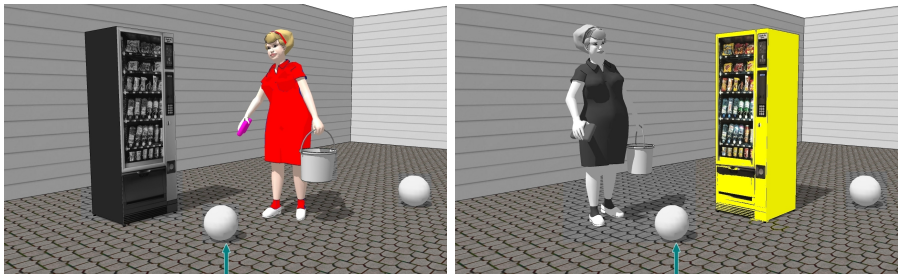
Fifty five native Dutch undergraduates from Tilburg University participated in this study for partial course credits (32 women, mean age 22 years). One participant was discarded for never mentioning a relatum.

Materials

Stimuli from Experiment 1b were used. In addition, one relatum candidate in each picture had a unique color (red, blue, green or yellow), while all the other were greyscale (see Figure 3).

Procedure

As in Experiment 1a. The position of the colored relatum candidate was counterbalanced across presentation lists.



(a) (b)

Figure 2.3: Experimental stimulus with on the right of the target object in color (red) the animate object (a) and in color (yellow) inanimate object (b)

Results and Discussion

Participants used one of the two possible formulations (43 participants mentioned both relata, 4 participants mentioned a single relatum) or produced mixed descriptions across trials with both single-relatum and two-relata formulations (7 participants). Due to small sample sizes, participants that opted for a single-relatum were grouped with those who used a mixed formulation and analysed as a mixed formulation group.

Whether spatial position influenced reference production was tested by comparing the chance of mentioning the left item as first relatum with random chance, using one–sample t–test. The chance of mentioning the left entity as first relatum was 61 percent (two-sided 95% CI [0.55, 0.66], $SD = 0.20$, $t(53) = 3.81$, $p < .001$, $d = 0.47$).

Whether animacy or perceptual salience overruled the left bias was analysed with an ANOVA test, having the Position of the Animate (2 levels: animate left, animate right) and the Position of the Coloured entity (2 levels: colored left, colored right) as within subjects factors, and Participant Formulation Preference (2 levels: two-relata, mixed) as a between subjects factor.

There was no statistically significant effect of the Position of the Coloured entity ($F < 1$).

There was a main effect of the Position of the Animate ($F(1, 52) = 18.645$, $p = .001$, $\eta_p^2 = .264$). Participants mentioned the left entity as relatum more often when the animate entity was placed on the right of the scene ($M = .67$) than when the animate was placed on the left ($M = .43$).

There was a main effect of Participant Formulation Preference ($F(1, 52) = 6.613$, $p = .01$, $\eta_p^2 = .113$). Participants mentioned the left entity as first relatum more often within a two-relata formulation ($M = .63$), than within a mixed one ($M = .47$).

There was an interaction between the Position of the Animate and Participant Formulation Preference ($F(1, 52) = 4.183$, $p < .05$, $\eta_p^2 = .074$). Speakers that used a two-relata formulation, mentioned the left entity as first relatum more often when the animate was on the right ($M = .70$) than on the left ($M = .57$). The same pattern of results was observed for speakers that used a mixed formulation (animate right $M = .65$, animate left $M = .29$). A split analysis showed that the general behaviour of the two formulation groups is essentially the same, but the effect size is higher for the mixed formulation ($F(1, 10) = 7.101$, $p = .024$, $\eta_p^2 = .415$), than for the two-relata one ($F(1, 42) = 7.809$, $p = .008$, $\eta_p^2 = .157$).

Experiment 1d revealed that perceptual salience, namely entities with unique colors, did not influence reference production, while conceptual salience had a small influence.

Experiment 1 has examined the extent to which the production of spatial relational descriptions is influenced by spatial position and salience of potential relata. Our results showed that spatial position indeed influenced reference production: relatum objects positioned on the left in the scene were more likely to be mentioned as

(first) relatum than those positioned on the right. However, participants did not systematically opt for the leftmost relatum object, suggesting that there might be other factors that could influence reference production as well. Therefore, in Experiment 1b - 1d, we manipulated the (conceptual and perceptual) salience of relatum objects, and these manipulations had no effect. In particular, we did not find that relatum objects that were salient, because of animacy, by priming visual attention or by using salient colors, were more likely to be used as (first) relatum. In Experiment 2, we assess if spatial position and salience affect listeners' evaluations of spatial descriptions.

2.3 Experiment 2 - Listener preferences

To further investigate the extent to which spatial position and salience might influence listeners' preferences for relata, in Experiment 2, participants were asked to rank relational descriptions. Given that many earlier studies have revealed strong effects of animacy, we expect descriptions that have an animate entity as (first) relatum to be ranked higher.

For pragmatic reasons, the language used in Experiment 2 was English. Earlier work on reference production (Koolen, Krahmer, & Theune, 2012; Theune, Koolen, & Krahmer, 2010) suggested that English and Dutch are comparable in terms of the attributes used in descriptions.

2.3.1 Participants

Eighty-six English-speaking native participants from Australia, Canada and the UK were recruited via CrowdFlower, a crowdsourcing service similar to Amazon Mechanical Turk. The validity of this method for behavioural studies has been previously tested and studies assessing data quality have been positive about using crowdsourcing as an alternative to more traditional approaches of participant recruitment (e.g., Buhrmester, Kwang, & Gosling, 2011; Crump, McDonnell, & Gureckis, 2013)). Ten participants' data were excluded for various reasons: because their ranking was identical (in more than 30 percent of the cases) to the order in which descriptions were presented (2 participants); because they declared being not native English speakers (5 participants); because did not finish the task (3 participants). The final sample included 66 participants (37 males, mean age 39.36 years, range 20 – 64 years).

2.3.2 Materials

The stimuli from Experiment 1b were used. The 32 experimental stimuli were divided across 6 randomized lists. The experiment consisted of 8 experimental stimuli (out of which 4 had an animate positioned left and 4 had an animate positioned right) and 8 filler stimuli. In addition, we used a set of four sentences representing the

two participant formulation preferences using a single relatum and two relata. These sentences were translated from Dutch to English. The sentences were: *the ball in front of the ANIMATE* (e.g., the man); *the ball in front of the INANIMATE* (e.g., closet); *the ball between the ANIMATE and the INANIMATE*; *the ball between the INANIMATE and the ANIMATE.*

2.3.3 Procedure

First, participants were instructed to rank the four descriptions starting with the one they "liked best" given the visual scene. The descriptions were presented under each scene in random order. The participant could rank the descriptions by dragging them in an input field with four empty slots, where the slot no. 1 represented the description that participants liked most, while slot no. 4 was assigned for the description that they liked least. The picture remained on the screen until the participants had made their choice and pressed a button to continue. Each experimental trial was followed by one filler trial.

2.3.4 Results and Discussion

For each trial, the order of the descriptions was ranked, starting from 1 (the best description) to 4 (the worst description).

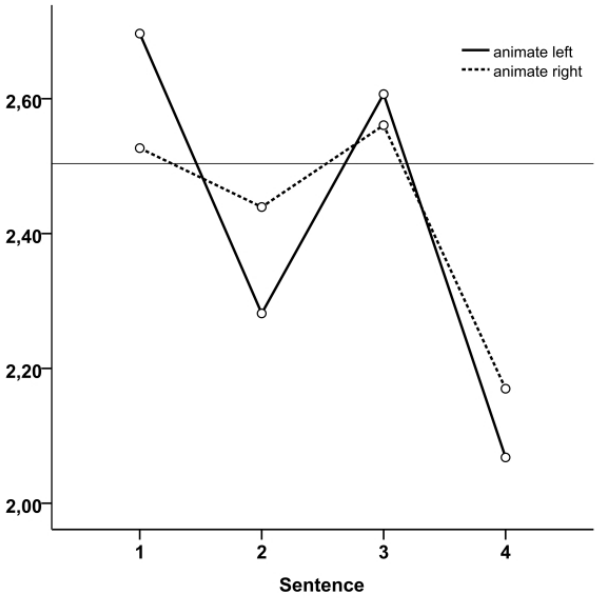
Whether animacy influenced preferences was tested with a repeated measures ANOVA, having three within subjects factors: the Position of the Animate (2 levels: animate left, animate right), the Participant Formulation Preference (4 levels: in front of ANIMATE, in front of INANIMATE, between the ANIMATE and the INANIMATE, between the INANIMATE and the ANIMATE) and Scenes (4 levels)[2].

Results revealed a main effect of Participant Formulation Preference ($F(3, 306) = 5.186$, $p = .002$, $\eta_p^2 = .048$) and a significant interaction between Animate Position and Participant Formulation Preference ($F(3, 306) = 4.412$, $p = .005$, $\eta_p^2 = .041$). Participants preferred the description that mentioned two relata and started with the animate irrespective of the visual scene (animate left $M = 2.07$, $SE = .11$; animate right $M = 2.17$ $SE = .11$) (see Figure 4). The second most preferred description was the one that mentioned a single relatum, namely the animate. This description was more preferred when the animate was positioned on the left of the scene ($M = 2.28$, $SE = .08$) than on the right of the scene ($M = 2.44$, $SE = .09$; $F(1, 102) = 6.58$, $p = .003$, $\eta_p^2 = .082$). The least preferred description was the one mentioning a single inanimate relatum, especially when the animate was placed on the left ($M = 2.70$, $SE = .09$; animate placed right $M = 2.53$, $SE = .09$; $F(1, 102) = 9.08$, $p = .012$, $\eta_p^2 = .061$).

[2]The analyses were also done using non-parametric Friedman's signed rank tests which yielded similar results.

2.4 Conclusions and Discussion

The main aim of this chapter was to examine the extent to which production of spatial relational descriptions is influenced by spatial position and salience. Our results show that spatial position systematically influenced reference production. A basic language elicitation task determined that speakers often mentioned the entity positioned leftmost in the scene as (first) relatum. This was consistent across four production experiments (highest mean 67 percent, η_p^2 range 0.47 – 0.75). Based on these observations, we considered that other factors might influence reference production. Thus, we investigated possible effects of the objects' (conceptual and perceptual) salience. In Experiment 1b, conceptual salience was manipulated visually, by having an animate and an inanimate relatum candidate. Despite the strong body of research arguing for effects of animacy in reference production, animacy was found to have a significant ef-



sentence 1: X in front of the INANIMATE

sentence 2: X in front of the ANIMATE

sentence 3: X between the INANIMATE and the ANIMATE

sentence 4: X between the ANIMATE and the INANIMATE

Figure 2.4: Mean ranks across conditions (1 = highest preference, 4 = lowest preference), where 2.5 represents random chance

fect in only one out of three production studies (Experiment 1d). Visual salience was manipulated using two different methods. In Experiment 1c, attention was primed using a flash and in Experiment 1d, the objects were made perceptually salient by having a distinctive colour. These manipulations yielded no effects. From a listener's perspective, the formulation of the description and the position of the animate entity in the scene influenced to some extent the acceptability rating (Experiment 2). These results are further discussed in relation to broader aspects of reference production.

2.4.1 Relevance for reference production

The studies reported in this chapter bring evidence for relatum reference being influenced by the inherent spatial structure of the scene, a factor largely unexplored in studies of (computational) reference production. Across different circumstances, there was a systematic preference for mentioning left entities as (first) relatum in relational descriptions such as *in front of X; in between X and Y*. This preference could have been caused either by cultural differences or spatial asymmetries in scene scanning. It is worth replicating Experiment 1 with speakers of a language with a right-to-left system.

The position of the object seems to be a constant factor influencing reference production. Our results are consistent with Miller et al. (2011), who stress that the spatial relation between the target and the relatum candidates is an important predictor in relatum selection. Congruent evidence comes from Clarke, Coco, and Keller (2013), who report that position (measured in relation to the centre of the screen) contributes to perceptual salience of the object and affects the likelihood with which objects are mentioned. When objects are symmetrically arranged, not only spatial position, but also salience influence (to some extent) referential choices.

Previous research has granted an important role to salience in reference production. Visually salient and linguistically important (e.g., animate) objects are more likely to be mentioned, as well as objects spatially placed in a prominent position (Clarke, Coco, & Keller, 2013). In these studies, we have manipulated salience on conceptual and visual levels. We expected salient entities to influence the ordering of linguistic elements in the spatial relation and be mentioned (first) more often than the other candidates. Surprisingly, there were poor effects of animacy, no effects of the visual salience manipulation. Below we address a few questions related to these results.

First, why did animacy have a limited influence on reference? The impact of animacy on word order, and more precisely on conjunctive phrases is debatable (see Branigan et al., 2008). For example, when the conjoined NPs are presented embedded in a sentence such as *the dog and the telephone were making noise* or *the surgeon yelled for a nurse and a needle* (experiments 1 and 2 in McDonald et al., 1993), animacy had no reliable effect on conjunct order. However, when removed from sentences and produced in isolated phrases (experiments 3, 4, and 5 in McDonald et al., 1993),

animate nouns regularly occupied a leading position. It is conceivable that the effect of animacy in the current studies might have been dampened by sentence context, in line with the findings of McDonald et al. (1993). Compared to other experiments that found a strong effect of animacy on reference production in visual domains (e.g., Coco & Keller, 2009), in our studies animacy was manipulated visually, without priming participants with animacy in a lexical format. 'Visual animacy' was suggested to be a less important factor in attention guiding (Wolfe & Horowitz, 2004). Interestingly, the results of the acceptability rating task (Experiment 2) present a different picture, which is more in line with previous studies suggesting strong effects of animacy and is in apparent contrast with the production data from Experiment 1. Descriptions which included an animate entity as the first (or the only) relatum were rated higher than those having an inanimate as first or single relatum. In fact, the descriptions which had animate as first relatum were rated as the most acceptable, irrespective of the spatial placement of the objects in the scene. Not only animacy, but also the left bias seemed to have influenced the acceptability ratings, as descriptions containing a single animate relatum, were rated higher when the animate entity was placed on the left, rather than on the right side of the visual scene and the same pattern was observed for descriptions that included a single inanimate relatum. This slight discrepancy between the results of Experiment 1 and 2 highlights an observation that has been made before in the context of REG evaluation: what speakers do is not necessarily what is appreciated most by addressees (for a review, see Krahmer & van Deemter, 2012; Gatt & Belz, 2010).

Second, why did priming attention have no effect? Directing speakers' attention to a specific region of the scene predicts which entity would be mentioned first, both in sentences and in conjoined NP descriptions (Gleitman et al., 2007). Yet, in our study, the attention capture cue did not influence utterances. Preference for left entities was stable, even when visual attention was directed to a different relatum candidate. It might be the case that the effect of the cue fades during production (the first-mentioned entity in our scenario was always the target ball). Other studies also report no effect of this attention priming manipulation (Arnold & Lao, 2015; Nappa & Arnold, 2014). In addition, when salience was explicitly manipulated by making an object perceptually salient, it did not yield a significant effect. This might be caused by the visual simplicity of the stimuli.

The extent to which our results can be observed using complex visual scenes also warrants further study. For example, Viethen and Dale (2008) reported (limited) effects of relatum salience in scenes consisting of three objects with simple spatial arrangements, but in a more complex study, salient large relata did not systematically influence whether the object was mentioned or not (Viethen et al., 2011). Similarly, participants describing routes through groups of colored objects in a MapTask (Louwerse et al., 2007) seem to have disregarded potential visual distractors (Viethen & Dale, 2011). The results of Elsner et al. (2014); Clarke, Elsner, and Rohde (2013) reveal a different picture: in very cluttered and complex scenes, like the Where's

Wally pictures, speakers were sensitive to perceptual salience, not only when choosing the objects to mention, but also when producing a description. The relational descriptions started more often with the salient object. Nonetheless, our studies are complementary, showing (though to a smaller extent) effects of the position an object occupies in the scene and salience.

Our experiments have a number of limitations. As mentioned above the scenes used as stimuli were simple and consisted of a small number of objects. Ideally, future research should take into account scenes of a higher visual complexity, use a different spatial arrangement of the objects and manipulate other perceptual features (such as size and orientation) as well. For a systematic analysis, other tasks should be considered as well (e.g., testing listeners' comprehension in a reaction time study).

In the production experiment, we also discouraged participants from saying "the ball on the left". While objects in visual environments can be referred to with a wide variety of forms of spatial language, we wanted to focus on referential choices when describing objects in relations. However, we also acknowledge that identifying a target by mentioning its location (and thus, maybe contrasting the target with a potential distractor, see Tenbrink, 2005) is a widespread strategy. Crucially, more research is needed to find out when people need or prefer relational descriptions containing explicit relata.

2.4.2 Formulation preferences

As for the formulations used, across studies, a small sample of participants chose a single relatum, thus producing a *X in front of Y* description. The chance of choosing one of the entities was not influenced by the distance between the relatum and the position of the distractor (the further away the relatum object was from the distractor ball, the less ambiguous).

Most of the participants referred to the target using the preposition *tussen* (in between), which describes the location of the target in relation to both relata. Compared with other locative prepositions, *in between* is a syntactically complex and cognitively more expensive one (because it contains more words and involves more relata), but it also provides a more accurate description. This preposition might be preferred due to the view point from which the speaker looks at the scene (Kelleher, Ross, Mac Namee, & Sloan, 2010), from which the relatum candidates and the target seem arranged in an almost linear fashion. In fact, when the target object is situated between two other elements and the in between relation is available for reference, speakers will often use this option (Tenbrink, 2007, p.261).

2.4.3 Recommendations for referring expressions algorithms

Understanding the criteria on which humans base their referential choices offers insights for the development of referring expressions generation algorithms. There are

only few algorithms that make use of extrinsic attributes as a last resort (e.g., Dale & Haddock, 1991; Gardent, 2002; Krahmer & Theune, 2002; Krahmer et al., 2003; Varges, 2005). Crucially, more research is needed to find out when people need or prefer relational descriptions containing explicit relata. Nevertheless, these systems have little to say about relatum reference as they assume access to a predefined scene model, where the relata has been selected and treat spatial reference as the last means for generating a unique description. Though there are some assumptions regarding the factors that drive choices regarding relatum reference, there is no systematic research on this issue. For example, Krahmer and Theune (2002) note that human speakers and hearers might have a preference for relata which are close to the target. Kelleher et al. (2005) implement a measure for proximity and bring into discussion visual and discourse salience. Dos Santos Silva and Paraboni (2015) consider distance as the main factor, followed by the unique spatial relations between objects. Apart from distance, various other factors may influence relatum reference. For example, Elsner et al. (2014) highlight that visual features that contribute to the object's perceptual salience should be taken into account in order to generate more human-like reference in visual domains. Specifically, perceptual salience (spatial and visual information) influences the order in which relata are mentioned in relational descriptions.

Our results suggest that algorithms should take into account the spatial position and the object's salience. When the distance between target and the relatum candidates is similar, the spatial structure of the scene should be the first feature to be examined. In circumstances in which there are several relatum candidates similarly aligned, we suggest that entities placed on the left of the target to be favored. Perceptual and conceptual salience might also be taken into account. Given the practical nature of REG, the human-likeness aspect should be balanced with a comprehension-oriented perspective (e.g., Garoufi, 2013; Paraboni, van Deemter, & Masthoff, 2007; Mast, Couto Vale, & Falomir, 2014). Our results suggest that if the goal of the system is different from just producing a human-like expression, other factors might play a role (see also Krahmer & van Deemter, 2012)). More addressee oriented (and maybe more efficient) descriptions might be produced when including an animate as first relatum. Our results suggest that when the target object is situated between two other objects and the *in between* relation is available for reference, the system should refer to both objects and start with the animate irrespective of the position of the objects in the scene. However, if the system generates a description with a single relatum, this relatum preferably should be the object located on the left of the target.

Finally, speakers have to make several referential choices when uttering spatial descriptions and different factors can influence this process. Our results suggest that reference production was affected by the spatial position of a relatum candidate and less so by (conceptual and perceptual) salience.

CHAPTER 3

Producing referring expressions in identification tasks and route directions: what's the difference?

Abstract Though communicative purposes are an important element in language production, few studies investigate the extent to which they might affect referential choices. In this chapter, we contrast two tasks with different purposes: identification and route directions giving. In Experiment 1, speakers referred to a target building nearby or further away, so that their addressee would distinguish it between other buildings (identification) or give route directions and use the same building as a landmark (instructions). Our results showed that irrespective of the speaker's purposes, referring expressions consisted of the same types of attributes, yet the attribute frequency and formulation differed. In the identification task, the referring expressions were longer, contained more locative and more post-nominal modifiers. In addition, referential choices were influenced by the visual distance between the speaker and the target: when speakers observed the target from far, their references were longer and contained more often locative modifiers. In Experiment 2, a different group of participants had to evaluate references produced in Experiment 1, while assessing descriptions of objects or descriptions of objects extracted from route directions. Neither task, distance, nor the length of the phrases influenced their choice, indicating that addressees consider references produced in both conditions equally adequate in both uses.

This chapter is based on: Baltaretu, A., Krahmer, E., & Maes, A. (2016). Producing referring expressions in identification tasks and route directions: what's the difference? *under review*

3.1 Introduction

Suppose you want to point out a building to a tourist, either because that is the hotel he is looking for or because it is part of the route direction you are asked to give. In both cases, you would have to describe the building (the target) in such a way that your addressee can distinguish it from the other buildings (the distractors). Most probably you will have to choose between different attributes (modifiers) that single out the building (e.g., color, location, size, architectural style). Though you need to refer to the same building, the two situations are rather different. In the first case, your primary purpose is to help your addressee distinguish the target from similar objects (e.g., "look at X"). This is similar to an identification (or discrimination) task where the speaker has to utter a distinguishing description in order to enable the addressee to identify the target object (van Deemter et al., 2012). Compared to an identification task, where the focus is on describing the target, when giving instructions, the distinguishing description is part of an action oriented speech act, with the purpose of reorienting the addressee on the correct street (e.g., "go to X and turn left"). The question that arises is to what extent humans tune their referential choices when having different communicative purposes (e.g., object identification, route directions)?

This chapter focuses on referring expression production and comprehension in naturalistic scenes, e.g., how a speaker chooses among different attributes, how this choice is influenced by communicative purposes and how addressees evaluate such references. We define route directions as a procedural (action oriented) discourse aimed at helping a person navigate in an unknown environment (e.g., Allen, 2000; Michon & Denis, 2001). This is composed of navigation actions (e.g., go, turn) specified by directions (e.g., left, right, straight) and descriptive information, such as reference points or landmarks (e.g., objects, such as buildings, and their attributes) (Allen, 2000). Paired with actions (e.g., "turn at X"), landmarks ground the direction change that has to be performed at an intersection (Michon & Denis, 2001) and have been shown to positively affect the quality of the instructions and the navigation performance (Tom & Tversky, 2012; Denis et al., 2014). By a discrimination / identification task, we refer to a situation in which the speaker has to produce a unique referring expression for identification purposes only. The referring expression is required to be a description consisting of a unique set of properties that singles out a target from similar, distractor objects (e.g., Koolen et al., 2011; van Deemter et al., 2012). The focus of this chapter is on initial (definite) references whose content is shaped by information available in the visual physical context. In general, these target descriptions consist of a definite article, a head noun, and one or more modifiers. Their production has been studied extensively in recent years (most notably in the Referring Expression Generation community), but typically in these studies the task context was not taken into account, which raises the question to what extent findings from one study, based

on references produced in response to a particular task, carry over to other contexts.

3.1.1 Communicative purposes and referring expressions

Communicative purposes are an important element of language production. There are reasons to believe that different purposes would trigger differences in the formulation of the referring expressions and differences regarding referential choices. Theoretically, an interaction includes the listener's and the speaker's assumptions of the communicative purpose (also referred as communicative intention cf. Levelt, 1993). According to Levelt's model of language production, in the conceptualization stage, speaker selects information ('what to say') and the linguistic ordering of this content within a sentence. The speaker would select and order information in such a manner as to satisfy the communicative purpose (Levelt, 1993, pp. 108 - 109). Similarly, the Maxim of Quantity states that speakers should "make a contribution as informative as required (*for the current purposes of the exchange*)" (Grice, 1975, p. 45; our emphasis). In other words, the maxim should be interpreted in a contextually sensitive manner, where the level of information depends on the communicative purposes of the speaker (Mooney, 2004). The nature of these purposes introduces a specific perspective of the situation (such as describing or instructing), which could influence the level of informativeness of a contribution (see for example Clark, 1996's work on dialogue).

A speaker can assess the situation with regard to the purpose that needs to be satisfied, and could shape references accordingly. The observation that the purposes of the interaction influence language production has been often acknowledged by qualitative work on dialogue development and some experimental work on reference and spatial language (Clark, 1996; Di Eugenio, Jordan, Thomason, & Moore, 2000; Daniel & Denis, 2004; Vorwerg & Tenbrink, 2007). Yet, as far as we know, there are no systematic comparisons assessing how much different referring expressions could be. We propose to investigate the extent to which different purposes affect reference production and comprehension, in a study that uses naturalistic scenes (depicting buildings in intersections), while taking into account a perceptual factor present in natural settings (the distance from which the target object is being observed). We compare the referring expressions produced when 'identifying' buildings, with the ones produced for the same objects while giving 'route directions', and assess how these references are evaluated by addressees.

There are several experimental studies emphasizing identification as a purpose in itself (for a review, see van Deemter et al., 2012; Krahmer & van Deemter, 2012). These 'identification studies' focus on the properties of the target in contrast to other objects, often presented in simple (sometimes grid-like) visual contexts (e.g., the GRE3D3 and the GRE3D7 corpora, Viethen & Dale, 2008, 2011; the TUNA and the D-TUNA corpora, van Deemter, van der Sluis, & Gatt, 2006; Koolen et al., 2011; the STARS corpus, Paraboni et al., 2016). The speaker's purpose is to identify a

target in such a way that the addressee picks out the target object with a particular description. Speakers mostly achieve their goal, but this ability is dependent on their attention to potential distractors (e.g. Brown-Schmidt & Tanenhaus, 2006), and sometimes they provide more information than is necessary (overspecification, e.g., Koolen et al., 2011). In general, identification is assumed to be part of a larger cooperative goal-directed interaction (e.g., Clark & Wilkes-Gibbs, 1986; Brennan & Clark, 1996). However, in the type of studies mentioned above, this interaction is rarely modeled and other purposes, apart from identification, are not taken into account. Some identification studies have explicitly addressed the Maxim of Quantity and investigated factors that might influence the level of informativeness of a conversational contribution (e.g., Engelhardt, Bailey, & Ferreira, 2006; Koolen et al., 2011; Arts et al., 2011), but these, too ignore the latter part of the maxim.

A recent tendency to move towards more natural tasks and complex scenes in identification studies has led to new resources. There are several corpora that have a broader purpose, namely instruction giving in the context of collaborative treasure hunting in 3–D virtual worlds (e.g., QUAKE, Byron & Fosler-Lussier, 2006; GIVE-2, Gargett, Garoufi, Koller, & Striegnitz, 2010; SCARE, Stoia, Shockley, Byron, & Fosler-Lussier, 2008). Here, data consist of whole conversations between partners cooperating on a task, making it difficult to isolate the impact of prior discourse context on the referring expressions used. Other corpora, such as ReferIt (Kazemzadeh et al., 2014) and REAL (Gkatzia, Rieser, Bartie, & Mackaness, 2015), consist of collections of real-world scenes and descriptions of a large sample of objects (ReferIt) and buildings (REAL) meant to help an addressee find the correct object among others. Gkatzia et al. (2015) argue that real world spatial scenes present a certain degree of perceptual ambiguity (e.g., object properties are less well defined, targets might be less visible if observed from further away), and references are syntactically and semantically much more complex (e.g., "the Austrian looking house, white house with the dark wooden beams at the water side" (Gkatzia et al., 2015, p. 1937; see also Clarke, Elsner, & Rohde, 2013). In fact, the amount of detail (e.g., the number of words, the length and the complexity of the referring expressions) positively influence addressees' performance, leading more often to correct object identification (Gkatzia et al., 2015). Though using naturalistic scenes, the ReferIt and REAL corpora also focus solely on identification. This raises the question to what extent the results of identification studies would generalize when speakers have different purposes. Do speakers produce similar references when describing an object for their addressee and when giving route directions, or do they adapt their references to the communicative context? How would addressees evaluate these referring expressions in a route directions vs. object description context, and how would the perceptual complex nature of the scenes influence reference production and evaluation?

Earlier work indeed suggests that different purposes might affect the attributes used in the referring expressions. For example, when the participants' goal was to negotiate, rather than identify, the preference for different attributes changed (Di Eu-

genio et al., 2000). In the COCONUT corpus, players had to buy items on a fixed budget, and in order to carry out this task, they had to describe and negotiate the furniture items that they believed were relevant to the task. In doing so, the item's price became one of the most often used attribute. Moreover, different purposes may result in referring expressions with different levels of specification. When speakers had to give an instruction (ask an addressee to physically move objects, e.g., 'Can you move the (small) plate to the left?'), they avoided scalar adjectives that were unnecessary for identification from a listener point of view. Contrastingly, when they had to describe events (inform the addressee, e.g., 'The experimenter will move the plate to the left'), they referred to the size of the object more frequently (Yoon, Koh, & Brown-Schmidt, 2012).

Not only communicative purposes, but also the importance attached to them might influence the speaker's reference production. When speakers had to identify 'buttons' on a control panel used by a surgeon, compared to identifying 'elements' for another participant to click on, they were more likely to include detailed descriptions containing redundant attributes and location information (Arts et al., 2011). It was also found that these helped the addressee fulfil the task faster. Similarly, when the participants' purpose was instructing someone how to operate once an alarm clock, compared to teaching someone that needs remember how to do this operation every night, their referring expressions contained more detailed information (Maes, Arts, & Noordman, 2004).

Taking stock: the production and comprehension of referring expression has received considerable attention, but earlier studies differed in the tasks that were given to participants (ranging from mere identification to, for example, negotiation) and the stimuli used to elicit references (ranging from a grid of line drawings to complex photographs). However, we still poorly understand the influence of these settings, since earlier studies never systematically compared tasks and stimuli to gauge their impact on reference production. This is unfortunate, because it means that it is unclear to what extent results from different studies are comparable and to what extent results obtained in one study carry over to a situation that is different from that particular study. To find out, in this chapter we report on a new study in which we systematically compare references produced in an identification task to those produced in a route directions giving task. Additionally, we include an extra factor in the design, the distance from which the participants perceive the target. We propose a study that uses naturalistic visual scenes (snapshots of Google Street View intersections), which are similar to natural settings with respect to the level of detail and complexity. In such situations, perception and recognition of object properties might be harder to assess for both speaker and listener. For example, the visibility of some object (or of its parts) may not be inferred with complete certainty, and a larger distance would affect the size of the target, the amount of visual details, and the number of objects in the visual field. Distance differences seem to trigger various strategies when producing references. Speakers tend to point more often when close to the object and refer

less to the target's location, but use more locative phrases when pointing becomes ambiguous (van Der Sluis & Krahmer, 2004; Bangerter, 2004).

3.1.2 The current study

In this chapter, we analyse whether referring expression production and evaluation is influenced by the purposes of the interaction (identifying and giving route directions). Our approach is similar to identification studies: we present participants with naturalistic scenes and we ask them to refer to a target building. Using the same set of visual scenes which depict buildings closer or further away from the viewer, we directly compare the extent to which the two aforementioned purposes and visual distance affect the form and content of referring expressions (e.g., length, types of attributes, their frequency and distribution). First, we elicit descriptions in two contexts, one in which identification is the sole communicative aim, and a second one where speakers need to give route directions and refer to landmarks (Experiment 1). Second, we evaluate preference for references produced in the two tasks, by asking participants to choose the phrases they prefer given a scene, keeping in mind that they are evaluating route directions or object descriptions (Experiment 2).

It is conceivable that references might have a different level of specification and the types of attributes and their distribution might be different across the two tasks. Different purposes might bring into attention different aspects, and a distinct focus of attention could influence the particular choice of features to be included in a referring expression (Beun & Cremers, 1998). For example, in the identification task, speakers have to refer to the target in contrast with the other buildings and they might mention more details about it (such as the number of windows on a façade). Longer, more detailed references could also be expected when there is a large difference regarding the distance between observer and the target which might increase the level of uncertainty. In such situation, when precision is important and the risk of misunderstanding is high, we would expect higher levels of informativity (e.g., more detailed, longer referring expressions containing more attributes) (Davies & Katsos, 2009, 2013). This would result in longer references, containing more attributes. When the intersection is far, the references could be longer, though the opposite could also be the case, since from nearby one has access to more visual details.

Contrastively, route directions have a strong instructional focus and procedural information conveying the turning direction is crucial for the success of the task. Though, referring to a landmark also requires an 'identification' step, it might be more important what the speaker is trying to do with the utterance (e.g., Searle, 1969), namely signal the correct road to be followed. As speakers need to convey two types of information (about both the building and the street), we might expect shorter, more focused references and maybe less locatives in the route directions (similar to the instruction giving GIVE-2 corpus, see Gkatzia et al., 2015).

Lastly, we might expect addressees to be sensitive to differences regarding the formulation of the referring expressions and the attribute choice. When evaluating route directions they might prefer the phrases previously produced in the route directions task, rather than those produced in the identification task. In line with Gkatzia et al. (2015), we could also expect that, irrespective of the communicative goal, addressees might prefer more complex, detailed references, especially when targets are further away and harder to perceive.

3.2 Experiment 1 - Production

3.2.1 Methods

Participants

Eighty native Dutch-speaking students (40 dyads) of Tilburg University (62 women, mean age 21 years) participated in exchange for partial course credits. Participants were randomly assigned to speaker roles (28 women). The study was carried out in accordance with the recommendations of APA guidelines for conducting experiments, and all participants gave written consent for use of their data.

Materials

Experimental materials consisted of 36 target objects (buildings depicted among other buildings in Google Street View snapshots of intersections). These targets were pictured from two distance points; 36 scenes taken with a camera positioned at 40 m away from the target (far condition) and 36 scenes taken with a camera positioned at 20 m away from the target (near condition). This resulted in 72 experimental scenes.

The target objects were marked with red squares, and placed in the corners of 4-way intersections (see Figure 1). These were always placed in the corners with highest visibility, namely on the other side the intersection. The targets' position was counterbalanced, so that in half of the scenes they were placed on the left side of the intersection. In the route directions task, we used exactly the same set of scenes, to which we added arrows indicating the the turning direction (see Figure 2). In addition, 36 filler scenes were added in order to present participants with a range of different navigation scenarios, and avoid participants from relying on fixed responses. The filler scenes depicted buildings in intersections with complex geometric structures.

Procedure

Participants worked in pairs, and completed their task on separate computers. Because we are interested in word choice and attribute use, pointing was discouraged by placing the computer screens in between the participants. The speaker received a

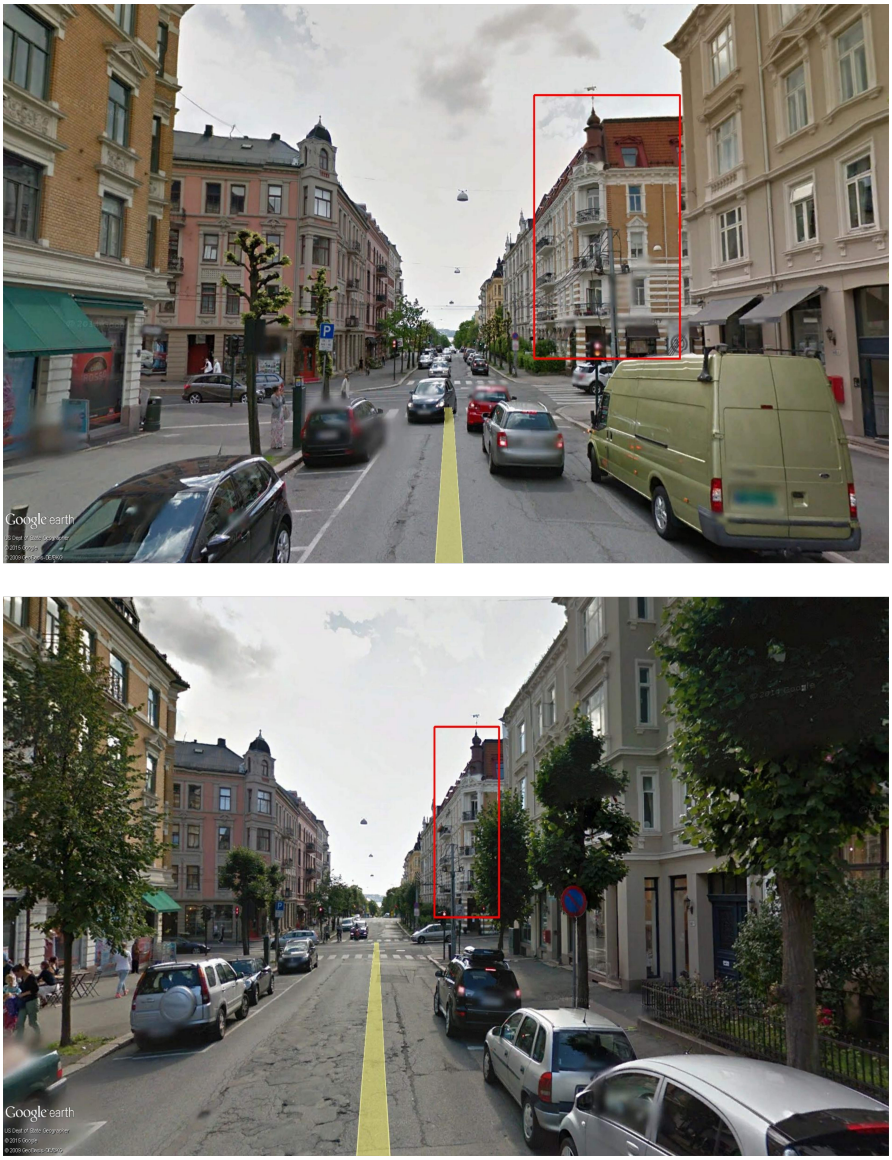


Figure 3.1: Experimental scenes from the description task depicting a target building near (above) and far (below)

scene with a marked building (and an arrow indicating the route), while the listener received the same image without any markings.

Participants were randomly assigned to either identify the target (a descriptions task) or identify the target while giving route directions. In the descriptions task (D-task), speakers were asked to refer to the marked building in such a way that the listener could uniquely identify it. In the route directions task (RD-task), speakers were asked to give route directions, and in doing so make use of the marked building as landmark. In order to elicit uniform responses, speakers were asked to verbally fill in templates (a typical procedure in identification studies, e.g., Dale & Viethen, 2009). They had to fill in the following templates: "click on ..." (descriptions) and "go to ... and turn left / right" (route directions). In both situations, the listeners had to click on the correct building. Listeners could ask questions if the speakers' instructions were unclear. Experimental scenes were divided in presentation lists (two lists for the D-task, 4 lists for the RD-task), so that each participant would see each target object only once. Participants were randomly assigned to one of the presentation lists. The experiment started with three practice trials, next 72 trials (36 experimental trials and 36 filler trials) were presented in different random orders.

Design and statistical analysis

This study had a 2 × 2 design with Task (levels: D-task, RD-task) as between participants factor and Distance (levels: far, near) as within participants factor. The dependent factors were the length of the references (number of words), the type of attributes mentioned in relation with the target (e.g., colour, location, etc.), the attributes' frequency and distribution (pre / post nominal modifiers), as well as the length of the remaining phrase after the speaker has mentioned the noun denoting the target building (number of words). In addition, we analysed lexical fillers / hedges which might be a signal of uncertainty in the speakers' instructions, and the addressee's error rates and questions, as an indicator of task difficulty. The presence / absence of attributes coupled with the target noun was binary coded. A phrase like "the large building on the left, next to the white tower" consists of a target noun ("building") and the following attributes: size ("large") and location ("on the left, next to the white building"). In order to analyse the differences regarding the references' length, data transformations (log data) were applied due to a skewed distribution. For ease of understanding, means and standard deviations reported here represent untransformed data.

In order to test the observed differences, we conducted separate statistical analyses using linear mixed model analysis (Jaeger, 2008), following the recommendations of Barr, Levy, Scheepers, and Tily (2013). We used the mixed logit model analysis as it can correctly account for random subject and item effects in a one-step analysis. The models were fitted using the LMER function from the LanguageR Package in R, version 2.15.2 (Bates, Mächler, Bolker, & Walker, 2015). To determine whether

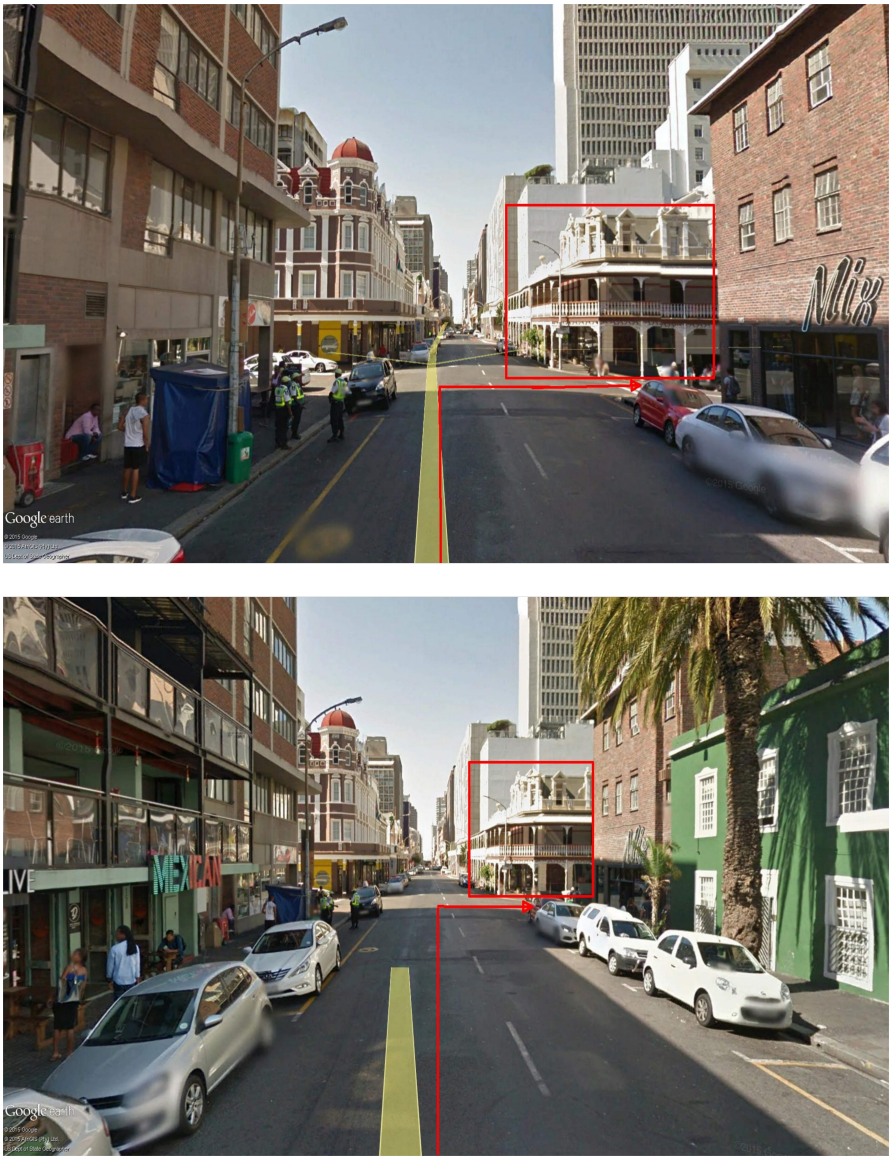


Figure 3.2: Experimental scenes from the route direction task depicting a target building near (above) and far (below)

the two conditions significantly differed from each other, we started by constructing a maximal model with a full random effect structure. This had Task and Distance as fixed factors; Speakers and Scenes as random factors; intercepts and random slopes for Speakers and Scenes to account for between-subject and between-item variation. When the dependent variable was binary coded, the factors were centered to avoid collinearity. In case the model did not converge, we only excluded random slopes with the lowest variance until convergence was reached. The results from the first converging model, as well as the structure of the model were reported. The p -– values were estimated via parametric bootstrapping over 100 iterations.

3.3 Results and Discussion

In total 1440 references were produced (40 speakers * 36 scenes). Data from two dyads (one from each condition) were discarded on the basis of task misunderstanding. The referring expressions consisted of a noun denoting the target object and all the phrases attached to it. In practice, in the D-task, the references consist of all the phrases after "click on", and in the RD-task, they consist of all the phrases between "Go to" and "and turn". In some cases, participants omitted some part of the template, but kept the overall structure (e.g., "The red building, and turn ..."), and these cases were included in the analysis (although "and turn" was obviously not included in the counts). In fact, 91 percent of the references had a similar structure (e.g., click on / go to + definite noun & modifiers + and turn direction, in route directions). The remaining, nine percent of the cases resulted in utterances with different structure (e.g., "The building is ..."), evenly distributed across the two tasks (5% cases in the D-task and 4% in the RD-task). As this structure could bias aspects such as word counts, we decided to exclude these cases from the analysis.

3.3.1 Length of referring expressions

The first converging model had random intercepts for Speakers and Scenes, R^2 marginal = 0.10, R^2 conditional = 0.56. The length of the references was significantly influenced by the Task ($\beta = -0.369$; $SE = 0.12$; $p < .01$). Referring expressions were longer in the D-task ($M = 16.3$, $SD = 1.3$), than in the RD-task ($M = 11$, $SD = 1.3$). There was a main effect of Distance ($\beta = 0.076$; $SE = 0.03$; $p < .01$). When close to the target, speakers produced slightly shorter referring expressions ($M = 12.8$, $SD = .88$), than when further away ($M = 14.03$, $SD = 1.01$). There was no interaction between Distance and Task ($p > .05$).

Type	Examples	Frequency	
		Route Directions	Descriptions
location	The second building on the left	68.0%	92.0%
color	The white building	42.0%	47.0%
building part	The building with (five) balconies / with (red) roof / with (two) windows	35.0%	36.0%
decoration	The building with stripes / with flowers pots / with hanging things / with flags	11.0%	12.0%
size	The smallest building	8.0%	7.0%
shape	The long building	0.6%	0.3%
age	The modern building	1.0%	0.5%
architectural style	The Italian building	2.0%	0.3%
materials	The brick building	1.0%	1.0%
evaluative	The ugly building	1.0%	2.0%

Table 3.1: Type of attributes, examples and attribute frequency split by task

3.3.2 Type of attributes

Speakers described targets by referring to ten types of attributes (see Table 1). The same types were produced in both tasks. Top three most frequent attributes in both tasks are location, followed by colour and references to structural parts of the target (such as chimneys, stairs, doors).

In the D-task, speakers mentioned more often the location and colour of the object than in the RD-task. We analysed the location difference statistically. The first converging model had random intercepts for Speakers and Scenes. There was a significant difference between the tasks regarding locative information ($\beta = -2.573$, $SE = 0.81$, $p < .01$). There was a main effect of Distance ($\beta = 0.535$, $SE = 0.22$, $p < 0.01$). When close to the target, speakers referred less often to the position of the object ($M = .77$, $SD = 0.41$), than when further away ($M = .82$, $SD = 0.32$). There was no significant interaction between Task and Distance ($p > .05$). Location was used to describe the target in almost all pictures (see Figure 3).

As for the difference regarding colour references, the first converging model had random intercepts for Speakers and Scenes, and there were no main effects ($p > .05$) or interactions ($p > .05$).

3.3.3 Distribution of modifiers

First, there was significant difference regarding the number of words produced after the target noun; the first converging model had random intercepts for Speakers and Scenes ($\beta = 0.508$, $SE = 0.16$, $p < .01$). The number of words produced after the

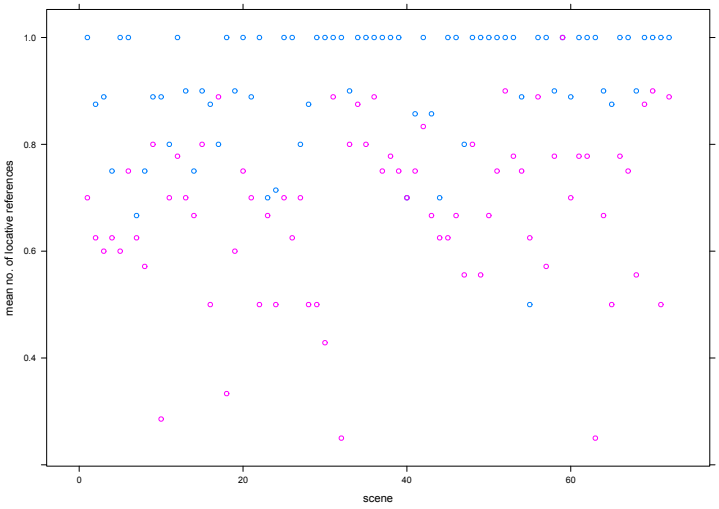


Figure 3.3: Frequency of locative information across scenes (distance: 'far' scenes 1 - 32, and 'near' scenes 33 - 72) split by task (D-task data in blue, and RD-task data in red.

	Pre-nominal		Post-nominal	
	Route Directions	Descriptions	Route directions	Descriptions
location	151	116	270	434
colour	128	70	136	221
size	21	17	27	28
other	6	9	250	319

Table 3.2: Distribution (number of cases) of pre- and post- nominal modifiers split by task.

target noun was longer in the D-task ($M = 13.62$, $SD = 1.25$), than in the RD-task ($M = 7.81$, $SD = 1.25$). There was no main effect of Distance ($p > .05$) and no interaction between the main factors ($p > .05$).

In general, there were more post-nominal modifiers in the D-task ($N = 1002$) compared to the RD-task ($N = 683$). The reversed pattern was observed for pre-nominal modifiers, which were more frequent in the RD-task ($N = 306$) than in the D-task ($N = 212$). In both tasks, pre-nominal modifiers consisted of location (e.g., "the left building"; "the second building"), colour (e.g., "the white building") and size references (e.g., "the large building") (see Table 2).

Post-nominal modifiers consisted also of location, colour and size, but mostly included references to structural parts of the building that syntactically can not be framed otherwise (e.g., the building with two balconies). Individual differences regarding the production of pre- and post- nominal modifiers can be observed in Figure 4.

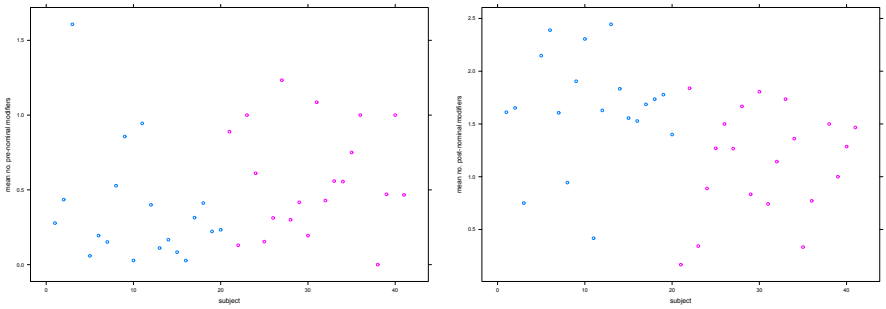
3.3.4 Lexical Fillers (hedges)

We define lexical fillers (hedges) as words or phrases that are conventionally used for signalling hesitation, marking the reference as more provisional (Brennan & Clark, 1996). The initial references produced by speakers in the two tasks included a different amount of lexical fillers (see Figure 5). D-task triggered more lexical fillers, compared to the RD-task.

3.3.5 Error rates

There was a small number of cases in which the addressee clicked on wrong buildings (3 cases in RD-task and 6 in the D-task) and relatively few clarification questions (13 questions in the D-task and 26 questions in the RD-task). With these questions, addressees mostly asked for simple clarifications (e.g., "a big building?") to which speakers uttered short confirmations (e.g., "yes").

In sum, when speakers had to identify a building, as opposed to identifying while giving route directions, they formulated their references differently: there were longer



(a) Pre-nominal modifiers (b) Post-nominal modifiers

Figure 3.4: Individual variation in the production of modifiers split by task (D-task speakers in blue, and RD-task speakers in red).

phrases, with more post-nominal modifiers (such as references to other buildings and structural parts of the target) and more lexical fillers. However, the references produced in the two tasks were similar from a semantic point of view. In both tasks speakers made use of the same types of attributes, albeit with some differences regarding the frequency with which some of these attributes were mentioned. The location of the target was mentioned less in route directions than in descriptions (the only statistically significant difference regarding attributes frequency). The distance between speaker and target influenced to some degree both the length of the reference and the frequency with which location was mentioned. Given these differences, we wonder to what extent addressees have preference for one type of reference or another. In Experiment 2, a different group of participants was asked to evaluate some of the references produced in Experiment 1, knowing that they are evaluating descriptions of objects or descriptions extracted from route directions.

3.4 Experiment 2 - Evaluation

3.4.1 Participants

Thirty-five native Dutch-speaking students of Tilburg University (21 women, mean age 21 years and 3 months) participated in exchange for partial course credits. The study was carried out in accordance with the recommendations of APA guidelines for conducting experiments, and all participants gave written consent for the use of their data.

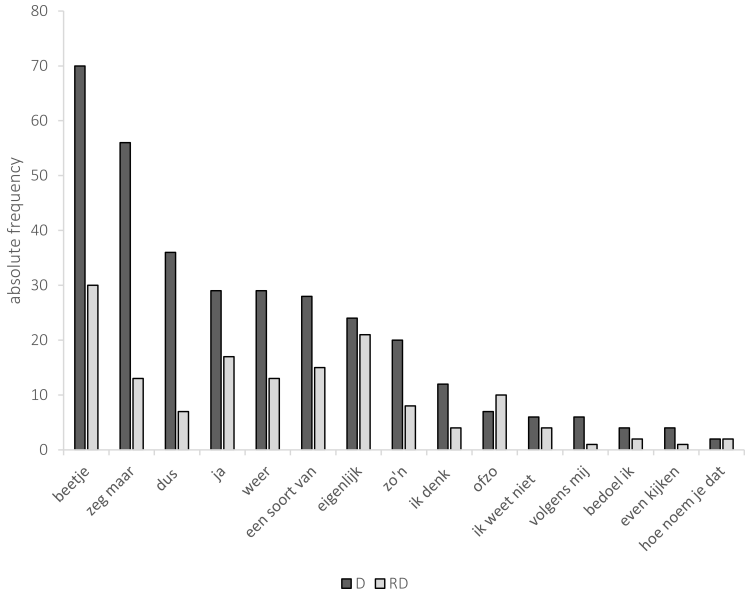


Figure 3.5: Distribution (number of cases) of lexical fillers split by task

3.4.2 Materials

The experimental scenes (72 scenes) consist of 36 target buildings depicted from far and from close distance, previously used in Experiment 1. In addition, a set of phrases describing the target were semi-randomly extracted from the corpus collected in Experiment 1 as follows. For each scene, we have chosen four phrases: two phrases produced in the RD-task (RD-phrases), and two phrases produced in the D-task (D-phrases). In total, 288 phrases were selected (72 scenes $\times$ 2 D-phrases $\times$ 2 RD-phrases). We made sure that the same phrase was not selected twice, and that the RD-phrases were not identical to the D-phrases. All the phrases started with a determiner ("the"), and were corrected for grammatical mistakes, repetitions, false starts and hesitations (see Figure 5 for a list of the words excluded).

3.4.3 Procedure

Participants were presented with 144 experimental trials (72 scenes shown twice in random order). Each scene was paired with two phrases (a D-phrase and a RD-phrase). The scene was presented on the upper three parts of the screen with the two phrases placed underneath (see Figure 6). The position on screen of the two types of phrases was counterbalanced and randomized, so that each type would be displayed an equal amount of times on the left and on the right side of the screen.

Participants' task was to judge the phrases content wise and choose (click on) the phrase that best described the target object. Participants were instructed that they were either evaluating descriptions of buildings (description evaluation condition, 'D-evaluation') or descriptions of buildings extracted from route directions (route directions evaluation condition 'RD-evaluation'). In the instructions, participants were shown an example of a description / route direction as produced in Experiment 1. The example consisted of a reference to the target building, highlighted with a different colour and embedded in the original template. The experiment began with two practice trials, followed by 144 experimental trials presented in random order. Once participants clicked on the phrase, the choice was recorded, and a new trial would automatically start. There were no time constraints.

3.4.4 Design and statistical analysis

This experiment had a 2 $\times$ 2 design with Evaluation Task (levels: RD-evaluation, D-evaluation) as a between participants factor and Distance (levels: far, near) as a within participants factor. The first dependent variable was the type of phrase chosen (a D-phrase or a RD-phrase) and second, if participants chose the longest out of the two phrases displayed together.

Statistical analysis was performed as in Experiment 1. The models included Evaluation Task and Distance as fixed factors, Subjects and Scenes as random factors,

Figure 3.6: Example of experimental scene presented together with a D-phrase (left) and a RD-phrase (right)

as well as random intercepts and random slopes for Subjects and Scenes. The first converging models and their structures are reported.

3.5 Results and Discussion

3.5.1 Type of phrase

The first converging model included random intercepts for Subjects and Scenes. There were no main effects (Evaluation Task $p > .05$, Distance $p > .05$) and no interactions between the main factors ($p > .05$).

3.5.2 Longest phrase

The first converging model included random intercepts for Subjects and Scenes. There were no main effects (Evaluation Task $p > .05$; Distance $p > .05$) and no interactions between the main factors ($p > .05$). Participants chose RD-phrases 52% of the time.

Participant's choice was not influenced by the type of evaluation task that they were performing or the distance at which the target was placed. Moreover, they did not have any preference for a specific type of phrase (short / long phrases or RD-phrases / D-phrases).

3.6 Conclusions and Discussion

In this chapter, we questioned to what extent different communicative purposes (object identification, giving route directions) influence reference production, while using complex naturalistic scenes and taking into account a perceptual factor, the visual distance from which a target object is observed. The referring expressions were elicited using real-world scenes, and the design allowed a direct comparison between the references elicited in two communicative settings. In Experiment 1, speakers had to describe a building for an addressee to distinguish it from other buildings or to give route directions and refer to the same building as landmark. In both conditions, the addressees had to click on the intended building. The referring expressions produced in Experiment 1 were then evaluated by a different group of participants (Experiment 2). These participants were presented with scenes coupled with two phrases (a description of an object and one of a landmark), and had to choose the phrase they preferred most. Data showed that identification as opposed to referring to objects in route directions triggered a number of differences regarding reference formulation, and almost no semantic differences.

First, there were no semantic differences between the two tasks. In Experiment 1, irrespective of purposes, speakers used the same types of attributes. In both tasks, location and colour were the most frequently mentioned attributes. In Experiment

2, we found further support for the semantic similarity of these phrases. When participants were asked to judge content wise which is the best description, they had no systematic preferences for a type of phrase, and being told explicitly that they were evaluating descriptions / route directions did not influence their choices. This pattern of results suggests that studies, where identification is the main purpose of the interaction (e.g., the TUNA corpus, van Deemter et al., 2006, the Where's Wally corpus, Clarke, Elsner, & Rohde, 2013), could generalize to a large extent to other settings, at least where the selection of semantic attributes is concerned. This is of particular importance for computational algorithms of reference generation, which are trained and tested on these corpora. These algorithms automatically convert data into text (Reiter & Dale, 2000), which is useful for instance, in the automatic generation of picture descriptions (Mitchell, van Deemter, & Reiter, 2010; Feng & Lapata, 2010) or navigation instructions (Garoufi & Koller, 2010). More recently these algorithms have been proven to be insightful models of human reference production and conceptualisation, because they make predictions that can be tested in psycholinguistic experiments (van Gompel, Gatt, Krahmer, & Deemter, 2012; Frank & Goodman, 2012).

However, the syntactic and lexical formulation of the references was different. Descriptions were longer and contained more post-nominal information and more words following the target noun, than references in route directions. This suggests that speakers described the target in more detail. Moreover, in the description task, speakers more often conveyed information about the location of the object. Locating an entity has been suggested to be a robust and successful strategy in object identification (Clarke, Elsner, & Rohde, 2015; Paraboni et al., 2016), which could have contributed to the increased length of the descriptions (see also Vorwerg & Tenbrink, 2007). Contrastively, in route directions, referring expressions were shorter and, to some extent, tended to contain more pre-nominal modifiers. Previous research suggests that attributes in pre-nominal position are more efficient for identification (Rubio-Fernández, 2016), which might suggest that when the purpose of the interaction is broader than just finding the correct object and the task has a higher level of complexity, speakers might 'optimize' their references, in such a way that the addressee can find the target faster and more easily (see also Clarke et al., 2015).

In addition, and unexpectedly, on the pragmatic level, we found different levels in the use of markers of nuance and hesitation. Lexical fillers add a degree of uncertainty, of non-commitment to a single way of characterizing an object, and they limit the truth condition of concepts (e.g., Lakoff, 1975). In naturalistic scenes, often buildings don't have a sharp attribute value, and speakers used different words to refer to the same property in order to characterize the object (e.g., "a red building, *but actually reddish*, *with a bit of* yellow"). Though it seems reasonable that the visual complexity of scenes (e.g., the lack of contrast) might trigger these self-corrections, they were more frequent in the description condition, suggesting that uncertainty is caused not only by what speakers see, but also by the task.

One may wonder whether the difference in absolute length between tasks was caused by the post-nominal structures, the pragmatic hesitations, or more words used to express some attributes, such as location. We conjecture it is a combination of these three. One possibility would be that nuanced references contain more modifiers, and the more words in modifiers, the more likely it is that they will be placed at the end of a noun phrase (e.g. in a relative clause following the target noun). Alternatively, the observed differences could be caused by the type of discourse elicited in the tasks. On the one hand, procedural information is of core importance for route directions. Speakers presumably want to make sure the addressee takes a turn on the correct street, and finding the correct building might be of secondary importance, which might lead to shorter landmark references. Moreover, landmarks are meant to improve route directions, and long complex references would doubtfully make an efficient contribution (for effects of conciseness on route directions, see Daniel & Denis, 2004). In addition, giving instructions is perhaps a better-defined task, where the identification of a building represents just half of the information to be conveyed, and sometimes in spontaneous direction giving, this reference might be missing altogether. In contrast, describing a target is more open-ended, and hence perhaps more difficult. Unless the addressee gives feedback and chooses the house, the speaker could continue incrementally adding attributes in an attempt to produce a more nuanced contribution. The higher amount of post-nominal information and lexical fillers could suggests that descriptions were less planned in advance.

Reference in real life is also characterized by visual contexts that are much larger and complicated than the ones typically employed in identification studies. We explored how a perceptual factor inherent to realistic settings, the distance between the speaker and the target, might influence referring expressions. Our results suggest that references were longer and contained more location information when the speaker was far from the target. However the differences between the two conditions were rather small (Experiment 1) and did not seem to affect the evaluation of references (Experiment 2). It would be interesting to explore the extent to which different aspects subsumed in the 'distance' factor contribute to perceptual uncertainty.

It should be noted that having other communicative tasks might have lead to different results. Different contexts in which the pressure of choosing the correct object might be higher than in route directions (e.g., instructing a nuclear power plant technician which buttons to press) would alter the level of detail of the descriptions. Maybe even having the same goal (e.g., identification), but formulating the task differently (e.g., describe clearly and thoroughly vs. describe creatively) could lead to differences in reference production. Moreover, psycholinguistic studies suggest that relying on some predefined ranking of attributes (for example location and colour should always be mentioned irrespective of task) cannot be applied straightforwardly in, for example, an interactive setting (e.g., Jordan & Walker, 2005), though it has been argued that in some contexts speakers use quick heuristics when selecting the content of their referring expressions (e.g., van Deemter et al., 2012).

In sum, when comparing the effects of different communicative goals on reference production an interesting pattern of results emerges. From a semantic point of view, irrespective of the purpose of the interaction, references consisted of the same type of attributes. There were several difference regarding formulation, with references produced in the description task being more nuanced. Despite formulation differences, when asked to judge the content of the phrases, participants did not have any preferences. We suggest that previous studies on identification can generalize to a large extent to other communicative settings.

CHAPTER 4

Improving route directions: the role of visual clutter and intersection type for spatial reference

Abstract In this chapter, we ask whether references to paths and landmarks in route directions (RDs) are influenced by environmental complexity, zooming in on two aspects of the visual surroundings, namely intersection structure and visual clutter. Speakers are asked to produce (Experiment 1), understand (Experiment 2) and evaluate (Experiment 3) turn-by-turn route directions in a naturalistic setting (Google Street View panoramic pictures). We find that increased levels of visual clutter and intersections with complex structures trigger more references to landmarks and paths when participants produce RDs, longer decision times to determine what the next correct step in a route is, and increased preference for landmarks. Finally, we discuss possible implications for automatic RD generation.

This chapter is based on: Baltaretu, A., Krahmer, E., Maes, A., (2015). Talking about relations: Factors influencing the production of relational descriptions. *Applied Cognitive Psychology, 29*(5), pp. 647 – 660. doi: 10.1002/acp.3145

4.1 Introduction

Imagine the following situation, probably not far in the future: you are somewhere in a city trying to find a café. You do not know how to get there, but you can use context aware pedestrian navigation software. Instead of finding your way using a paper or a small size phone map, you are able to listen to turn-by-turn instructions that include references to objects you see around. Yet, we know little how properties of the (visual) environment influences the turn-by-turn production and comprehension of RDs. Route directions assume a complex interaction between the perception of the environment and the navigation task. In this chapter, we focus on the former by varying the level of detail in visual scenes (henceforth visual clutter). We address the latter by manipulating the conceptual complexity of turning actions in intersections with different geometrical shapes (e.g., turning right in a +- shaped intersection compared to a **K**- shaped intersection, Klippel, Tenbrink, & Montello, 2013).

By visual clutter, we refer to the amount of visual information (the density of items) in a scene. Visual clutter might be a potential problem for RDs: in a crowded, visually noisy environment it might be more difficult to find the way and to give good instructions. Clutter represents the state, organization and representation of items in which visual search performance starts to degrade; and it can be quantified by using the feature congestion algorithm (Rosenholtz et al., 2007). We propose to investigate whether visual information affects reference production and comprehension in route directions, in a study that uses naturalistic scenes (depicting intersections from a route perspective), while controlling for a perceptual factor (visual clutter) and a task related factor (the intersection structure).

It has been theorized that RDs are influenced by the structure of the environment (Richter & Klippel, 2004). We focus on one aspect related to the complexity of the navigation layout (in addition to visual clutter): the structure of the intersections. When several branches in an intersection head into the same direction (as the right branches of a **K**- shaped intersection), speakers have a harder task to refer to a street (on the right) in an unambiguous fashion. This might influence the number of references to landmarks (three dimensional points of reference, such as buildings) and paths (streets) that people include in their instructions. Describing the right street can be considered a referential task similar to an object identification task (Krahmer & van Deemter, 2012). The speaker has to select and refer to object properties that are relevant given the visual context and produce a description that would help an addressee identify the street. Language production and scene perception are closely intertwined (e.g., Spivey, Tyler, Eberhard, & Tanenhaus, 2001). However, little research has been conducted on naturalistic scene perception in relation to reference production and comprehension. Similarly, in the field of spatial cognition, RDs have been elicited for different intersection structures depicted on schematic maps that do not incorporate the visual richness of the environment (e.g., Klippel et al., 2013).

The question is how this task-related factor would influence reference to streets in route-view naturalistic scenes, across both production and comprehension processes.

Understanding how people make use of the richness of the visual context and adapt their instructions to the characteristics of the environment provides important behavioural evidence for developing future automatic RD systems. While for humans it is natural and easy to produce and understand references to objects from the environment, for machines this is still a challenge (for a discussion, see Richter & Winter, 2014; van Deemter et al., 2012). There is a wide range of objects people could refer to (paths and landmarks) and also a variety of ways to refer to these objects. Enabling computers to use the same type of references and adapt to the context in the same way as human speakers do, could lead to a more natural and easier human-computer communication. Given the practical nature of RDs (navigation aids meant to help one find his way), the human-likeness aspect should be balanced with a comprehension-oriented perspective, in which the efficiency of various instructions is tested (e.g., Garoufi, 2013; Paraboni et al., 2007).

In this chapter, we discuss three experiments that study the influence of visual clutter and intersection structures on RDs discourse. First, we looked at the effects of intersection and visual clutter on speakers' RD production (Experiment 1). Next, we assessed the time addressee needed for processing instructions with / without landmarks and deciding on which street to turn, as a function of visual clutter and intersection structure, as well as the accuracy of their responses (Experiment 2). Lastly, we checked participants' preference for instructions with landmarks across scenes of various complexity (Experiment 3).

4.1.1 Path and landmark references

In this chapter, by RDs we refer to a set of instructions on how to (incrementally) follow a route (Richter & Klippel, 2004; Allen, 2000). This type of discourse triggers two kinds of communicative goals: instructing the user on how to go from one location to another (via instructions for actions) and descriptions of the environment (via referring expressions). Hence, RDs include an action prescription coupled with a direction (go left), and can be enhanced with references to the visual aspects of the environment: path information (first street) and landmarks (the pharmacy).

Landmarks are defined as environmental features that function as points of reference (Allen, 2000) and structure mental representations of space (Richter & Winter, 2014). In the literature, the term landmark has been used to refer to the path (mostly two-dimensional entities) and to entities which are not part of the path (mostly three-dimensional entities) (Westerbeek & Maes, 2013). In this chapter, we will use the term path reference for two-dimensional entities which are part of the route structure (e.g., road intersections, Klippel and Winter (2005); side streets, squares, Denis, Pazzaglia, Cornoldi, and Bertolo (1999)). Paths can be referred via definite descriptions (the first street) or proper names (Church Street) (Tom & Denis, 2004; Tom & Tversky,

2012). We will use landmark reference for three-dimensional entities that are positioned along the path (such as buildings, monuments, etc., Denis et al., 1999). Thus, a change of direction can be specified via a path reference go left on the first street, a landmark reference go left past the pharmacy or by referring to both go left on the first street, past the pharmacy.

Even though RDs are not limited to these types of information, references to landmarks are considered to be crucial for these instructions (see Denis et al., 1999; Lovelace et al., 1999; May, Ross, Bayer, & Tarkiainen, 2003; Richter & Klippel, 2004). Given that referring to landmarks, a key ingredient, could maximize the helpfulness of the RD (Allen, 2000), we ask when and why people refer to landmarks and to what extent landmarks help pedestrian navigation in environments of different complexity. First, we argue that in turn-by-turn RDs, landmarks would be mentioned when necessary, namely when path references are insufficient to distinguish unambiguously the intended street from its competitors. This situations might be modulated by environmental complexity. For example, in simple intersections target streets can be fairly easy referred to via a path reference (turn right on the next street) and even target streets in complex intersections can be disambiguated by using ordering concepts (e.g., in a **K**- shaped intersection turn right on the second street; for the robustness of this strategy, see Klippel et al., 2013). Thus, minimally, a turn-by-turn RD could include an action verb, a direction and a path reference. Few studies analysed the rationale of adding references when internal and external landmarks are equally available to use (Westerbeek & Maes, 2013) or the extent to which these different reference types influence addressee's performance or preference. Thus, we evaluate how efficient and attractive are detailed instructions.

Second, we suggest that in complex environments, references to landmarks could actually be more beneficial and help the addressee choose the correct street that (s)he needs to follow. In complicated intersections, people take longer to navigate and make more mistakes (Montello, 2005), while the level of visual clutter was shown to increase the time needed for finding an object in a visual scene (Asher, Tolhurst, Troscianko, & Gilchrist, 2013). In a complex situation, even if the target street can be successfully referred to a via path reference, mentioning landmarks could be truly helpful for the addressee.

4.1.2 Visual Clutter

Earlier studies have suggested that information that grabs visual attention influences RD production and comprehension (e.g., Sorrows & Hirtle, 1999). Visual attention is influenced by object-related properties, such as the number and organization of objects in a scene (or visual clutter). In a visually noisy intersection it is harder to see the correct street. How do people cope with increased visual complexity and what should a system providing turn-by-turn navigation instructions do in such situation? Should it adapt the instructions, and if so, how?

In earlier research, clutter has been studied in relation with visual search, map reading and linguistic processing. We briefly review these aspects, highlighting the possible influence of clutter on RDs comprehension and production. The excess of items and their disorganized display lead to crowding and occlusion, thus decreasing object recognition performance (Bravo & Farid, 2006), and increasing the difficulty of both segmenting a scene (Bravo & Farid, 2004) and performing visual search (Henderson et al., 2009; Neider & Zelinsky, 2011; Asher et al., 2013). The number of objects in a scene positively correlates with the reaction times for finding a target (Rosenholtz et al., 2007). The more visually complex a scene (more objects added to the display), the longer the visual search times and the poorer the visual search efficiency. In general, high levels of visual clutter have been shown to be detrimental to performance across different tasks. Cluttered displays affect map reading due to hiding or masking essential information (Agrawala & Stolte, 2001) and visual clutter is a good predictor for search time and search errors. For example, when participants were asked to find buildings in maps with a satellite view, search times and errors were fastest/smallest for sparsely cluttered rural scenes, and increased across more cluttered (sub-urban, urban) scenes (Neider & Zelinsky, 2011). The same pattern of behaviour was observed when participants had to find targets on marine and aerial radar displays (Donderi & McFadden, 2005). Previous research suggests that it might be difficult in more cluttered scenes to find the correct street or the relevant landmark, and thus participants might need more time to decide and make more mistakes when choosing a street.

Visually cluttered scenes have been shown to influence language production. For example, the more complex or detailed the visual environment, the longer it took respondents to start typing their response and resulted in the production of more complex constructions and in more complex eye tracking patterns (Coco & Keller, 2009). When referring to objects, the amount of visual variation increases the redundant use of attributes (Koolen et al., 2013). Moreover, in cluttered scenes, referring to a particular object lowers the chance of speakers comparing the object they need to describe to all the other objects (of the same type) present in the scene (Koolen et al., 2013). Instead, in order to successfully distinguish the intended object from the rest, participants refer to nearby salient objects via locative expressions (Clarke, Elsner, & Rohde, 2013). Congruent evidence comes from a memory based route production task where the number of landmarks is kept constant across conditions (Westerbeek & Maes, 2013). Participants who had previously seen cluttered scenes added landmarks more often in their RDs, than those who had seen maps with lower levels of visual detail. We hypothesise that RD production might be also influenced by this factor. We expect that when scenes are harder to process, participants would produce more details, namely more references to landmarks and longer instructions. For a speaker that has to unambiguously identify a target (in this case a street), visual crowding might reduce visibility and increase the difficulty of comparing the target

street against the others in the intersection. As an alternative, speakers would rely less on features of the target and refer more often to salient neighbouring objects (Clarke, Elsner, & Rohde, 2013) that offer a higher degree of referential determinacy (Allen, 2000).

4.1.3 Intersection Type

Apart from the number of objects present in the environment, RDs might become more detailed as the route is becoming more complex. RDs could vary depending on the structure (branches and angles) of the intersection and the action required. In simple intersections street branches are intersecting at 90 degrees angle, the number of turning options is limited and the level of uncertainty is low. In this type of intersections, we would expect instructions to include a minimum amount of information (e.g., reference to action, direction and path). The complexity of an intersection increases not only with the number of branches and their intersecting angles (structural complexity), but also with the number of turning options in the same direction (conceptual complexity, e.g., turn right in a **K**– shaped intersection, Klippel et al., 2013). All these aspects increase navigational uncertainty, which in turn might influence RDs production and comprehension. How can an automatic system providing RDs deal with this uncertainty?

Human speakers providing RDs have different strategies to cope with increased complexity. For example, they can draw the addressee's attention by explicitly announcing that there is a point of the route that might lead to confusion and refer frequently to visual details of the route (Hirtle, Richter, Srinivas, & Firth, 2010). Similarly, in a RD production task, Klippel et al. (2013) noticed that complex intersections triggered longer descriptions, with more references to the intersection's structure and more alternative instructions on how to proceed, than simple intersections. In order to refer unambiguously to the street on which to turn, participants used different strategies depending on the type of intersection. They were naming the structure (e.g., fork right), comparing the possibilities to take (e.g., furthest right), adding locative attributes (e.g., the third to your left), describing the competing directions not to take. Given that people seem to have different strategies to refer unambiguously to the target street via path references, it is an open question when speakers would refer to landmarks. Based on these observations, in complex intersections, we would expect people to produce detailed descriptions (more path references).

From an addressee's perspective, in both types of intersections, having explicit path references might suffice to find the correct street, but having landmark information might increase their efficiency and accuracy in choosing the correct street. Psycholinguistic studies on object identification suggest that descriptions of objects which incorporate more attributes than necessary can facilitate the search of a referent in complex spatial domains, when the participant is in difficult situation (lack of orientation or dead end). Such descriptions lowered search time and search distance

(Paraboni & van Deemter, 2014; Arts et al., 2011). Yet, it is unclear to what extent referring to landmarks in intersections with different structures is beneficial for somebody that needs to decide on which street to continue.

4.1.4 The current study

In this chapter, we address the following research questions: do visual clutter and intersection structure influence route directions and if so how? How beneficial is it to add landmarks to instructions when people have to decide which street to take in more or less complex environments? And which type of instructions (with or without landmarks) do they prefer? We report on the results of three experiments touching on both RD production and comprehension processes in a complex environment. We argue that the complexity of the visual context, given a particular turning action, influences (1) the type of references (path and landmarks) speakers include in the RDs; and (2) the efficiency and preference for these instructions for somebody who needs to decide on which street to continue walking. The dependent variables in our experiments are the number of references to paths and landmarks that speakers include in their RDs (Experiment 1); the time and accuracy of addressee's performance when presented with instructions enhanced (or not) with landmark references (Experiment 2) and addressee's preference for landmarks (Experiment 3).

We predict that the amount of visual clutter and intersection structure would influence the way speakers disambiguate the target street from potential distractors, such as the other streets present in an intersection. In scenes with high levels of clutter, participants might add more landmark references. In scenes with complex intersection structures, we predict that RDs would include more path and landmark references. In our experimental materials, we use scenes with a route view, and in each scene objects that could be used as potential landmarks are located around the intersections.

Next, the effectiveness of different types of instructions (with / without landmarks) was tested by asking addressees to decide on which street to continue; and preference for instructions with landmarks was assessed by asking participants to choose the RDs (with / without landmarks) they prefer for each scene. Good route directions should ease the process of route finding and presumably addressees should decide faster. We expect landmarks to be helpful and preferred in situations with high levels of uncertainty, namely in scenes with complex intersections and high levels of clutter.

4.2 Experiment 1 - Production

4.2.1 Methods

Participants

78 English native participants from Australia, Canada and the UK were paid to take part in the experiment via CrowdFlower, a crowdsourcing service similar to Amazon Mechanical Turk. The validity of this method for behavioural studies has been previously tested and studies assessing data quality have been positive about using crowdsourcing as an alternative to more traditional approaches of participant recruitment (e.g., Buhrmester et al., 2011; Crump et al., 2013). Data from 35 respondents were excluded from the analyses because they were not native English speakers, did not finish the task, or misunderstood the task. The final sample included 43 participants (15 males, mean age 44 years, range 17 – 69 years). None of them was born or living in the locations from where the scenes were taken.

Materials

A pool of approximately 200 scenes was created by taking snapshots of rural and urban intersections in Google Street View. These scenes depicted typical streets of rural (low levels of visual clutter) and urban (high levels of visual clutter) areas from Australia, France, Hong Kong, Japan, Romania, South Africa, Spain, United Arab Emirates, USA; none of these scenes included famous tourist landmarks (such as The Eiffel Tower) that could be recognized by the participants. Both 'Asian' and 'Western' scenes occurred in both high cluttered / low cluttered scenes and in high / low intersection complexity scenes. The angle from which the snapshot was taken gives as much as possible a pedestrian perspective over the streets. The street names and other information specific to Google Street View (exact address and map) were occluded.

Based on Klippel et al. (2013)'s taxonomy, two scene types were created: scenes containing intersections either with simple structures (**T**- and +- shaped) or with complex structures (**Y**- and **K**- shaped, as well as crossroads with 5 branches). Orthogonal to this, the level of visual clutter in these pictures was estimated using the Feature Congestion algorithm (Rosenholtz et al., 2007) and human ratings. First, visual clutter is defined by pixel local variability of colour, orientation and luminance contrast over the entire image (Rosenholtz et al., 2007). Scene clutter has been previously explained as the absolute number of items in the scene (Branigan et al., 2008; Koolen et al., 2013). However, other factors (e.g., item arrangement, shape, and colour) might also have a contribution. We choose Feature Congestion (a state-of-the-art algorithm which yields similar performance with other methods of measuring clutter), as it is the only model that also captures the contributions of colour variability. Based on the clutter scores, 60 scenes were selected from the pool: 32 pictures

with low levels of clutter (1.93 – 2.73 feature congestion), and 28 pictures with high levels of clutter (3.70 – 5.20 feature congestion). Second, these 60 pictures were evaluated by a different group of 26 human participants, who rated how cluttered they considered a scene to be. Participants were asked to judge on a 7–point scale the complexity of the pictures, ranging from simple, low visual noise to complex, high visual noise. Based on these ratings, we chose 36 stimuli items (see Figure 1 and Figure 2 for examples) in which we counterbalanced visual clutter (high clutter 9 scenes; low clutter 9 scenes) and intersection structure (complex intersection 9 scenes; simple intersections 9 scenes). Finally, yellow lines depicting the route and the direction to be followed (left, right and straight) were drawn using an open source editor. In complex intersections, the yellow line was drawn always in the direction where more than one option was possible (such as a right turn in a **K**- shaped intersection).

Procedure

The instructions for participants specified a scenario similar to the example we gave in the beginning of this chapter. It stated that we are developing software that can generate real time/live pedestrian route descriptions based on the visual input coming from the Google Glass video camera and realized in audio format via a smartphone [1]. The task was to provide route instructions in English for a fictitious addressee who had to take the path marked with the yellow line. Participants saw one scene at a time and filled in the RD in the input field provided under the picture. The task started with 3 warm-up trials, then 36 experimental trials were presented in random order. There were no time constraints. Lastly, they filled in a series of demographic questions. The participants were recruited online and participated voluntarily. The experiment was waived by the ethics committee at Tilburg University.

Design and statistical analysis

Experiment 1 had a within participants design with Intersection type (levels: simple, complex) and Clutter (levels: high, low) as independent variables. The dependent variables were path references (number of references to channels of movement), landmarks (number of references to visual objects), and description length (number of words). These RD components were analysed separately using logit mixed model analysis (Jaeger, 2008), with Clutter and Intersection type as fixed factors; participants and item scenes as random factors. Random intercepts and random slopes for participants and item scenes were included to account for between-subject and between-item variation. First, a model with a full random effect structure was constructed (Barr et al., 2013). If the model did not converge, we excluded random slopes with the lowest variance. For the converging model, model comparisons were used to remove random slopes that did not contribute to the fit of the model according to a

[1] the full instructions for all three experiments are given in the Annex at the end of this chapter

Figure 4.1: Simple and complex intersections in scenes with low level of clutter

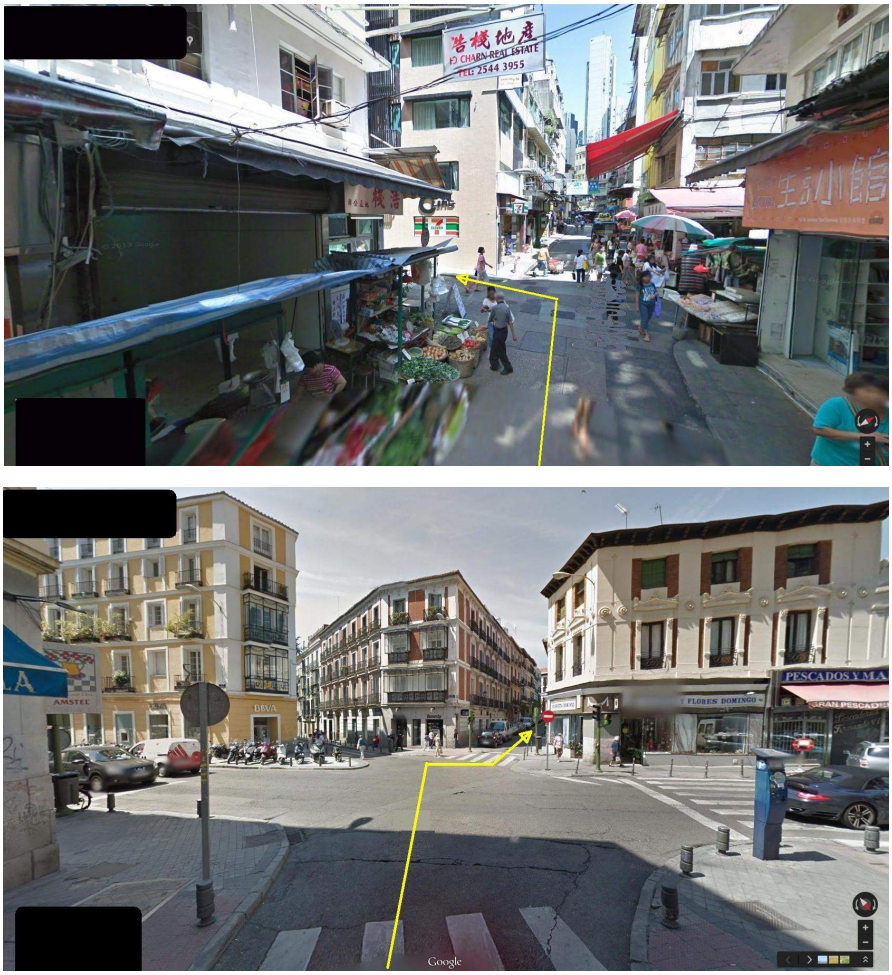


Figure 4.2: Simple and complex intersections in scenes with high level of clutter

	Low Clutter		High Clutter	
	Simple Int.	Complex Int.	Simple Int.	Complex Int.
Path	0.60 (0.56)	0.86 (0.76)	0.61 (0.61)	1.02 (0.78)
Landmark	0.11 (0.32)	0.13 (0.37)	0.17 (0.41)	0.33 (0.61)
No. words	5.40 (3.84)	7.23 (4.95)	5.81 (4.62)	8.03 (5.56)

Table 4.1: Means (standard deviations) of path references, landmark references, number of words for each clutter level split by intersection (int.) structure

likelihood ratio test. Only the final model is reported (p – values were estimated via parametric bootstrapping over 100 iterations).

4.2.2 Results and Discussion

There were 1548 RDs produced for this experiment. These contained 1208 path references and 227 landmark references (see also Table 1).

Path references

For the number of path references there was a main effect of Intersection type ($\beta = 0.386$, $SE = 0.16$, $p < .05$). Simple intersections trigger less path references ($M = 0.60, SD = 0.59$), than complex intersections ($M = 0.94, SD = 0.77$). Model $R^2 = 0.48$.

Landmark references

For the number of landmarks there was a main effect of Clutter ($\beta = 0.87$, $SE = 0.32$, $p < .001$). RDs in low cluttered scenes had fewer landmarks ($M = 0.12$, $SD = 0.35$) compared to scenes with high clutter levels ($M = 0.25$, $SD = 0.52$). In addition, there was a significant interaction ($\beta = 0.62$, $SE = 0.44$, $p < .05$) between the main factors: low cluttered scenes triggered in both types of intersection similar numbers of references (complex intersections $M = 0.13$; simple intersections $M = 0.11$). In high cluttered scenes there were more landmark references in complex intersections ($M = 0.33$) than in simple intersections ($M = 0.17$). Model $R^2 = 0.42$.

Length of descriptions

For the overall number of words in the route descriptions there was a main effect of Intersection type ($\beta = 0.286$, $SE = 0.12$, $p < .05$). RDs for simple intersections ($M = 5.61$, $SD = 4.25$) are shorter than those for complex intersections ($M = 7.63$, $SD = 5.28$). Model $R^2 = 0.65$.

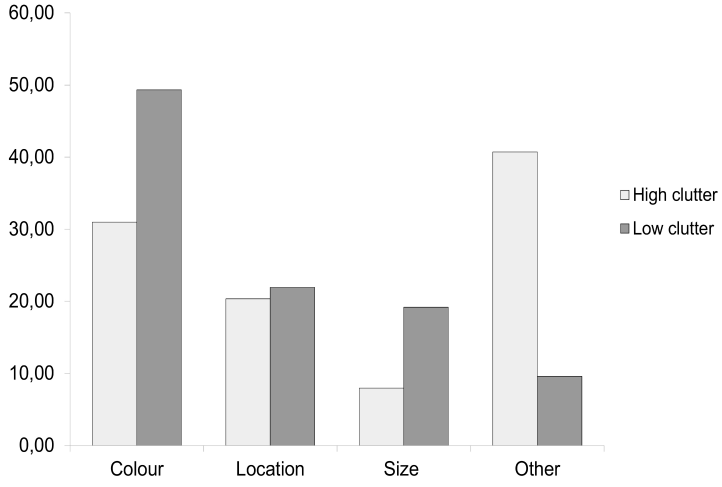


Figure 4.3: Distribution (percentages) of nominal modifiers in landmark references

Attributes used to refer to landmarks

In this experiment, there were no restrictions on what types of objects speakers could choose or on the content of these descriptions. Due to the fact that only visual clutter had a significant influence on landmark references, this analysis is focused on the two clutter conditions.

There were 227 references to 98 different landmarks. 76 references consisted of single nouns (e.g., "pharmacy") and 151 had different nominal modifiers such as colour ("red building"), size ("big building"); location ("the building on your left / at the corner of the street"), and other modifiers (brand names "Sushi Point" or materials "glass building"). Sometimes speakers referred to landmark parts, as in the yellow house with red roof. This construction was annotated as one landmark reference with two mentions of colour. Figure 3 shows the overall distribution of the modifiers across the two clutter conditions. In general, colour was the most often used attribute, followed by location and size. In the high clutter condition, these modifiers had a similar distribution, however, compared to the low clutter condition, they were mentioned less often. Contrastively, other types of modifiers were used. As the other category consisted mostly of brand names, this result might be due to the type of scenes: urban scenes displayed more often business names than rural scenes.

In addition, there were different types of landmarks mentioned. Given that we presented participants with static scenes, stable objects (e.g., buildings) were fre-

quently chosen: 77.6% landmarks (including trees and shop names which have a relative temporal stability), while 23% of the entities mentioned were atypical landmarks with a limited stability (parked or moving cars and pedestrians passing by).

Experiment 1 had participants with a wide age range and different backgrounds. Gender could not be added to regression models because we do not have enough data for a full analysis. However, inspection of the data revealed that gender did not seem to play any role.

Experiment 1 confirmed that the complexity of the scene affects the production of RDs, in length and in composition. Visual clutter significantly influence landmark references, while intersection structure influenced path references and the length of the instructions. In Experiment 2, we assess if these factors affect RDs comprehension. In addition, we test the effectiveness of two types of instructions (with / without landmark references) when participants need to decide the street on which to continue.

4.3 Experiment 2 - Comprehension

4.3.1 Methods

Participants

78 native Dutch students of Tilburg University (26 men, mean age 22, 1 years) participated in exchange for partial course credits. None of these participated in Experiment 1.

Materials

The stimulus materials consisted of 64 RDs and 32 scenes. The RDs were created as follows: a first set of 32 RDs were selected from the data collected in Experiment 1 so that (1) each instruction consisted of an action verb coupled with direction ("go right") and (2) had sufficient path information for making a correct choice (on the first street). The perspective employed was always of a pedestrian on route. These RDs have been translated in Dutch. Based on this first set, the second set of RDs was created ($N = 32$) by adding one landmark reference to each instruction. The landmark added to each RD was the most often referred object for that scene given its frequency in the corpus collected in Experiment 1.

The scene set consisted of 32 images used in Experiment 1 with several small changes as follows: a red arrow showed the position where the viewer is standing and the direction which he is facing; the streets were marked with 4 dots of different colours that corresponded to 4 keyboard keys marked with the same colours (see Figure 4 and Figure 5). These dots marked the four choices that participants could make based on the RD.

Procedure

The instructions specified the same scenario as in Experiment 1 and that now we are in a testing phase. The participants' task was to read the RDs and afterwards choose as fast as possible the street previously indicated. Participants saw 32 trials as follows: first a fixation cross was displayed for 500 ms, followed by a RD. When they finished reading (self-paced), participants could press any of the 4 keys to continue. Next, a scene would be displayed and participants had to decide which street they need to turn on according to the RD they read. To mark their decision, participants had to press the key that had the corresponding colour with the chosen street. After pressing the key, a new trial would start automatically. There were no time constraints imposed. Before the start of the experiment participants had time to accommodate and learn the position of the coloured keys and the experiment started with 4 warm-up trials followed by 32 randomized experimental trials. During the entire experiment, participants had to keep their index and middle fingers from both hands on the keys.

Design and statistical analysis

Crossing the factors Clutter (levels: low, high), Intersection type (levels: simple, complex) and Instruction type (levels: instructions with, without landmarks) resulted into a $2 \times 2 \times 2$ design. In the regression models, Clutter and Intersection type were included as within participants factors and Instruction type was included as a between participants factor. We measured reaction times (the time a participant needed to take a decision measured from the moment the picture was displayed until the participant pressed a key) and accuracy (if the participant made a correct decision or not). These were analysed separately using logit mixed model analysis with Clutter, Intersection type and Instruction type as fixed factors; participants and item pictures as random factors. Statistical analyses were done as in Experiment 1. Data transformations (log data) were applied on reaction times due to the skewed distribution. In addition, to better model the data, the statistical models used a poisson distribution. For the accuracy rate analysis the factors have been centred in order to avoid collinearity. For ease of understanding, means and standard deviations reported here represent untransformed data.

4.3.2 Results and Discussion

Reaction times

There was a significant effect of Intersection type ($\beta = 0.39$, $SE = 0.15$, $p < .05$). Participants decided faster which street to take in simple intersections ($M = 2363ms$, $SD = 1410$), than in complex intersections ($M = 3487ms$, $SD = 2173.26$), see also Table 2.

Figure 4.4: Simple and complex intersections in scenes with low level of clutter

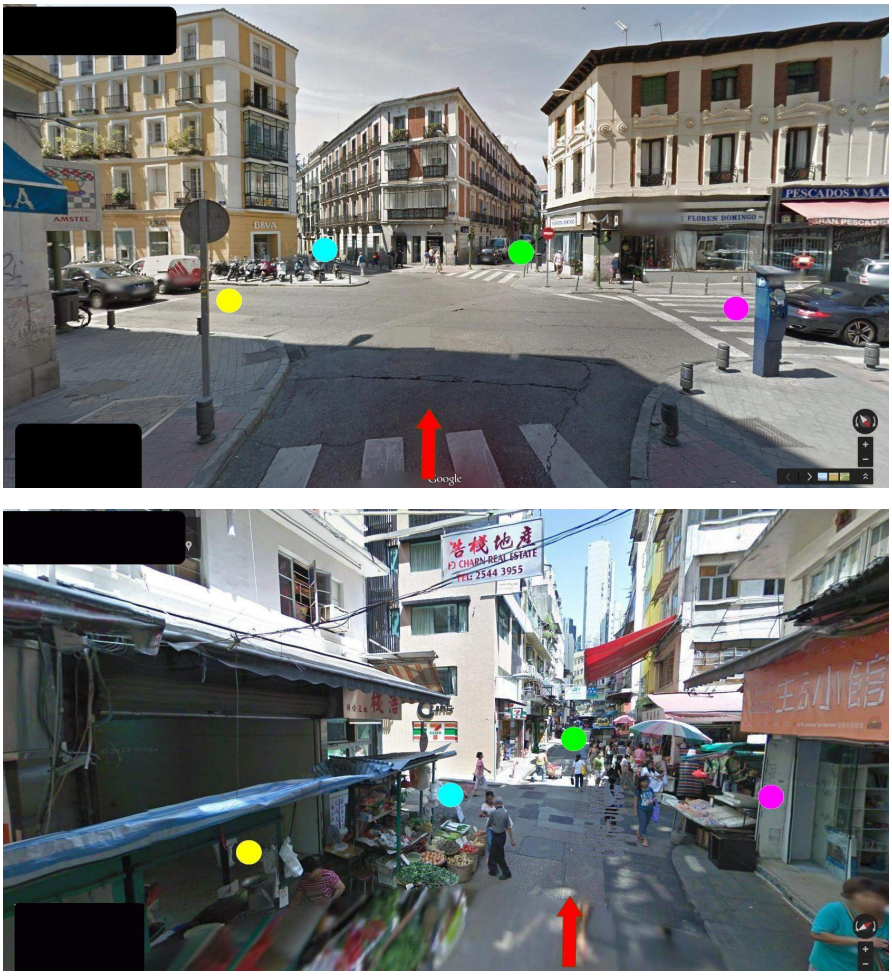


Figure 4.5: Simple and complex intersections in scenes with high level of clutter

Landmark	Low Clutter		High Clutter	
	Simple Int.	Complex Int.	Simple Int.	Complex Int.
With	2386 (1413)	3322 (1756)	3067 (1618)	4064 (2322)
Without	1931 (1191)	2716 (1485)	2020 (1032)	3821 (2689)

Table 4.2: Means (standard deviations) reaction times (ms) for instructions with and without landmarks; clutter condition split by intersection structure

Landmark	Low Clutter		High Clutter	
	Simple Int.	Complex Int.	Simple Int.	Complex Int.
With	0.91 (0.29)	0.93 (0.26)	0.91 (0.29)	0.86 (0.35)
Without	0.86 (0.35)	0.93 (0.26)	0.96 (0.20)	0.76 (0.43)

Table 4.3: Means (standard deviations) accuracy rates for instructions with and without landmarks; clutter condition split by intersection structure

There was an effect of Instruction type ($\beta = 0.25$, $SE = 0.08$, $p < .001$). Participants responded faster after reading instructions with no landmarks ($M = 2622ms$, $SD = 1883$), than instructions with landmarks ($M = 3210ms$, $SD = 1904$). This effect might be due to the different length of the instructions (instructions with landmarks were longer). Model $R^2 = 0.60$.

Accuracy rate

Overall, participants choose the correct route in 88.73% of the cases (see also Table 3). There were no main effects. There was a significant interaction between the Intersection type and Instruction type ($\beta = 1.31$, $SE = 0.71$, $p < .01$). While in simple intersections having or not landmarks did not influence the accuracy rate ($M = 0.91$, $SD = 0.29$ correct choices with landmarks; $M = 0.914$, $SD = 0.29$ correct choices no landmarks), in complex intersections participants gave more correct responses when receiving landmark information ($M = 0.89$, $SD = 0.31$ correct responses with landmarks; $M = 0.84$, $SD = 0.36$ correct responses with no landmarks).

These results suggest that the complexity of the environment influences to some degree the time needed to choose a street. In complex intersections participants needed more time to decide than in simple intersections, while the manipulation of visual clutter didn't reach significance. Regarding the efficiency of different types of instructions, those enhanced with landmarks, enabled participants to make more often correct decisions in complex intersections, but also elicited longer reaction times. Apart from participants' performance, we wanted to assess to what extent people perceive these instructions as being useful. Thus, in Experiment 3, we asked them to choose between the two types of instructions the one that best represented the visual

scene.

4.4 Experiment 3 - Evaluation

4.4.1 Methods

Participants

The same like in Experiment 2.

Materials

The stimulus materials consisted of 64 RDs from Experiment 2 and 32 scenes used in Experiment 1.

Procedure

The participants' task was to read the RDs and choose the one that they liked most. Participants saw 32 trials as follows: first a fixation cross was displayed for 500 ms, followed by a slide that consisted of a Google Street View scene (displayed on the upper 3 quarters of the screen) and both types of instructions (with / without landmark) placed below the scene. Participants had to choose one RD by clicking on it. Once the choice was recorded a new trial would automatically start. The experiment began with 4 practice trials, followed by the experimental trials presented in random order. The position of the two types of RDs on screen was counterbalanced. There were no time constraints.

Design and statistical analysis

Crossing the factors Clutter (levels: low, high), Intersection type (levels: simple, complex) and Group (levels: in Experiment 2 the participant received instructions with, without landmarks) resulted into a $2 \times 2 \times 2$ design with Clutter and Intersection type as within participant factors and Group as a between participant factor. We counted the number of times participants choose instructions with landmarks. The data was analysed using logit mixed model analysis with Clutter, Intersection type and Group as fixed factors; participants and item pictures as random factors. All factors were centred in order to avoid collinearity. Analyses were done as in Experiment 1.

4.4.2 Results

Out of 2496 cases (32 scenes x 78 participants), RDs with landmarks were chosen 647 times (26%).

Landmark	Low Clutter		High Clutter	
	Simple Int.	Complex Int.	Simple Int.	Complex Int.
With	0.17 (0.38)	0.30 (0.46)	0.24 (0.43)	0.60 (0.48)
Without	0.06 (0.23)	0.19 (0.40)	0.09 (0.29)	0.40 (0.49)

Table 4.4: Means (standard deviation) preference for instructions with landmarks; group split by clutter condition split by intersection structure

Preference for landmarks

There was a main effect of Clutter ($\beta = 1.18$, $SE = 0.35$, $p < .01$). RDs containing landmarks were chosen more often for high cluttered scenes ($M = 0.34$, $SD = 0.48$) than for low cluttered scenes ($M = 0.18$, $SD = 0.38$); see also Table 4 .

There was a main effect of Intersection type ($\beta = 2.03$, $SE = 0.38$, $p < .001$). Instructions with landmarks were chosen more often for complex intersections ($M = 0.38$, $SD = 0.49$), than for simple intersections ($M = 0.14$, $SD = 0.35$).

There was a main effect of Group ($\beta = 1.15$, $SE = 0.35$, $p < .01$). The participants who received in Experiment 2 instructions with landmarks preferred this type of instructions more ($M = 0.34$, $SD = 0.47$) than the other group ($M = 0.19$, $SD = 0.39$).

There was a significant interaction between Clutter and Group ($\beta = 0.52$, $SE = 0.24$, $p < .05$). RDs with landmarks were more often chosen for scenes with a high level of clutter by both groups (group without landmarks $M = 0.25$, $SD = 0.43$; group with landmarks $M = 0.44$, $SD = 0.50$), than for scenes with low level of clutter (group without landmarks $M = 0.13$, $SD = 0.33$; group with landmarks $M = 0.23$, $SD = 0.42$).

The participants' preference for instructions with landmarks was assessed. Overall, instructions with landmarks were not often chosen. This might be explained by the fact that the minimal instructions had always enough path information to make a correct decision. Nonetheless, instructions with landmarks were preferred in scenes of high environmental complexity. There were main effects of clutter and intersection type suggesting that participants chose more often instructions with landmarks when the scenes were cluttered and when the intersections had a complex structure.

We observed a priming effect: participants choose more often the type of instructions they were previously exposed to. However, irrespective of group, all participants showed a stronger preference for RDs with landmarks when the level of clutter in the scenes was high.

4.5 General Discussion and Conclusions

In this chapter, we studied the effect of environmental complexity, namely visual clutter and intersection structures, on RDs production and comprehension. Specifically, in Experiment 1, we analysed the extent to which these two factors influence references to paths and landmarks and the instructions' length. Both factors affected RDs in length and composition. Next, we assessed the efficiency of different types of instructions (with / without landmarks) across scenes of various complexity (Experiment 2), and the participants' preference for these RD types (Experiment 3). In complex environments, instructions with landmarks triggered more correct answers and were preferred more often. A detailed discussion of these results is given bellow.

In Experiment 1, 80% of the total number of references are path references suggesting that referring to paths is a basic strategy for identifying the target street and these references alone can be fully discriminatory, a result in line with Denis et al. (1999) and Westerbeek and Maes (2013). The amount of path references produced were influenced by the intersection structure, with complex intersections eliciting more references to paths and subsequently longer instructions. This might be due to the fact that the target street had to be singled out of a (higher number) of potential distractor branches. Not only might the number of branches (number of options in an intersection) lead to this effect, but also the complexity of the turning action (Klippel, 2003). However, because in this experiment the simple and complex intersection conditions consist of a variety of structures with different number of branches, we suggest a systematic analysis of this type for future research.

Participants also referred to landmarks and these references were mainly affected by visual clutter. Scenes with high levels of clutter triggered detailed descriptions with more references to three dimensional objects. In cluttered environments, the target street might be harder to distinguish and referring to paths (the most often used identification strategy) becomes suboptimal in differentiating between objects of the same type. This leads to a higher number of references to other objects in the proximity that could ground the turning action. Moreover, an interaction between visual clutter and intersection type suggests that speakers refer to landmarks especially in complex environments.

In addition, the types of landmarks and the content of the referring expressions were analysed. The objects varied in terms of stability: 23% of these entities were parked or moving cars and pedestrians. Given the existing literature on wayfinding, these results may seem surprising. Yet, we consider these objects natural points of reference for people in live situations. The reason why they are underrepresented in most standard navigation studies, is that the set-up of these studies often implies some kind of (temporal and / or spatial) asymmetry between the speaker and addressee perspective, thus making movable entities unreliable reference points. In this experiment, we synchronized the two perspectives. We found further evidence for this in a similar study presented in the next chapter of this thesis, in which we used

moving scenes as experimental materials, where the proportion of moving landmarks appeared to be even higher.

Speakers often referred to perceptual properties of landmarks, especially colour and location (e.g., "the yellow building on your right"). Determining the content of these descriptions becomes relevant given novel proposals of enriching datasets via crowdsourcing by asking users to describe landmarks (Richter & Winter, 2014, p.167). An open question is to what extent the content of these landmark references is similar given different tasks: describing objects in isolation vs. descriptions embedded in instructions, and if this affects the effectiveness of the RDs. In addition, in the cluttered urban scenes, landmarks were often referred to via brand names. Urban spaces display a wider range of businesses (and logos) on the streets and this explains why overall the use of colour, position and size is lower in the cluttered conditions. This could be a limitation of the stimuli, but also signals that some methods used by commercial systems to extract landmarks have a limited applicability in the rural and sub-urban spaces (e.g., Nokia City Scene system, in Wither, Au, Rischpater, & Grzeszczuk, 2013).

Given that in complex environments speakers refer more often to landmarks, we questioned to what extent these detailed instructions influence an addressee's decision times, as a function of visual clutter and intersection structure. In Experiment 2, we asked participants to read RDs with / without landmarks and decide after on which street they should continue. There was a significant effect of intersection structure on the time participants need to make a decision, with longer reaction times for complex intersections. We expected the same pattern of results for visual clutter, however, this factor did not reach significance. Previous research has shown that clutter influences visual search to a high extent. However, in this experiment the visual search task might have been influenced by the fact that the intersection' branches were marked with distinctive colourful dots. Marking the branches could have given addressees some advantages: the dots might have captured attention, and subsequently this might have reduced the amount of distractors to four possible options. Regarding the instruction type, participants who received instructions with landmarks were slower than those who received minimal instructions without landmarks. This delay might be caused by the fact that those who received both path and landmark information had to search two entities, and the instructions with landmarks were longer. Experiment 2 provides evidence regarding the circumstances in which referring to landmarks might be beneficial. Even though the task was relatively simple and the accuracy rate was overall high, there was a significant interaction between the intersection structure and the type of instructions. Only in situations where confusion might arise and navigational uncertainty was high (complex intersections), having received instructions with landmarks lead to higher accuracy rates.

In addition, we assessed the perceived effectiveness of these instructions. In Experiment 3, participants were asked to choose the instructions that best fit the scenes. Instructions with landmarks were overall chosen to a small degree. This

might be explained by the fact that all instructions had enough path information to make a correct choice. Despite the fact they were never necessary, references to landmarks were preferred in scenes with a higher degree of environmental complexity. Participants chose more often RDs with landmarks when the scenes were cluttered and the intersections complex. In addition, irrespective of the group to which participants were assigned in Experiment 2, all participants showed a stronger preference for RDs with landmarks when the level of clutter was high.

Finally, in turn-by-turn RDs, the street that needs to be followed has to be referred to in an unambiguous manner. Our findings have several implications for automatic RDs generation. We first suggest how the present findings could help a system adapt to the context and second, briefly discuss the content of these references.

4.5.1 Implications for automatic RD generation

Most automatic route generation machines have a natural language generation system responsible for the production of uniquely identifying descriptions of the target streets and potential landmarks (referring expression generation). Commercial systems are mostly limited to use references to street names, which are extracted from geo-databases. For several reasons, these instructions do not resemble human production of RDs (Tom & Denis, 2004; Tom & Tversky, 2012). People refer not only to streets, but also to three dimensional objects from the environment. Given state-of-the-art technology a system that has access to both the visual stream of information a person is exposed to and to the traditional street network database, could make informed decisions of when to add references to landmarks in the instructions. Our results suggest that in simple situations, such as simple intersections in a low cluttered environment, RDs that include (only) relevant path references would probably be sufficient (and preferred) by users. Navigation systems should add references to landmarks when the level of visual clutter is high, irrespective of the type of intersection. This idea is supported by the fact that (1) the number of landmark references is influenced by visual clutter, and (2) addressees judge these references as being useful, even when the instructions have sufficient path information for a correct decision. In addition, a system that makes use of landmarks in complex situations might reduce navigational uncertainty and help users make correct choices.

To wrap up, the three experiments reported in this chapter show that environmental complexity, more specifically visual clutter and intersection structure, influence RDs production and comprehension. References to paths are the most used strategy to refer unambiguously to the street that should be taken. Speakers refer to landmarks especially in urban environments with high levels of visual clutter. Additionally, addressees perform better and prefer receiving instructions with landmarks, but only in complex situations.

Annex: Instructions from the three experiments

4.5.2 Experiment 1

Dear participant, We are developing software that can produce real time / live spoken route directions for pedestrians. Imagine walking on the street with your smartphone and wearing a pair of special glasses ("Google glass"). These glasses have an integrated camera that sends direct visual input of where you are looking, to your smartphone. Thus, everything you see can be used by the app to help you find your way. At this stage, we are improving our software. Based on the visual input received from the camera, we want to add spoken route descriptions that you could listen to via your smarthphone while walking (see below).

At this stage, we are collecting good route descriptions that can be applied in the scenario described above. Therefore you are going to see pictures of street views containing a marked route line (see below).

The street views represent what a user of our pedestrian navigation system sees in real time. The pedestrian cannot see the yellow line. What you need to do: Consider yourself walking together with this person on the street. Take advantage of the fact that you both see everything that is on this street in the same time. You can refer to any aspect of the picture you wish. How would you instruct the person to walk in each scene? Please write down your route description in the designated input field. The input field is under the picture.

4.5.3 Experiment 2

Dear participant, We are developing software that can produce real time / live spoken route directions for pedestrians. Imagine walking on the street with your smartphone and wearing a pair of special glasses ("Google glass"). These glasses have an integrated camera that sends direct visual input of where you are looking, to your smartphone (see below).

Thus, everything you see can be used by the app to help you find your way. At this stage, we are testing our software. We want you to tell us how good these automatically produced route directions are. Each trial you will see two slides: one with a route direction the second one with an image

What you need to do: Read carefully the route direction. When you are ready press any coloured button and an image will appear.

The image shows the intersection that you have just read about. The red arrow marks the spot where you are and indicates the direction in which you were walking. The coloured dots show four different possibilities in which you can continue walking. One of these possibilities was described in the route directions.

The dots have the same colour in each picture and are always placed in the same order on the image, from left to right: yellow, blue, green, violet. These four colours

correspond in this order with the four coloured buttons on the keyboard. You must press the key corresponding to the described route.

Place the index and middle fingers on the four coloured buttons. Press as soon as possible on the key corresponding to the colour of the dot indicating the route described.

4.5.4 Experiment 3

You will see a picture of an intersection and two route directions. On each picture an arrow is marking the described route. Read the directions carefully. Click on the description that you like the best.

CHAPTER 5

Landmarks on the move. Producing and understanding references to moving landmarks

Abstract There is a general agreement that landmarks in route directions should be perceptually salient and stable objects. Yet, other attributes, such as (animated) motion, can also attract visual attention and make entities salient. In the present study, we investigate if and when speakers refer to moving entities in route directions and how listeners evaluate such instructions. We asked speakers to watch short videos of different crossroads with and without moving landmarks and give directions to listeners, who in turn had to choose a street on which to continue (Experiment 1) or choose the instruction they most preferred among three route directions (Experiment 2). Results reveal that speakers mentioned moving entities, especially when the trajectory was informative for the place where a turn should be taken (Experiment 1). Listeners had no problem understanding instructions with moving landmarks (Experiment 1). Yet, participants chose instructions with stable landmarks more often (Experiment 2). These results are discussed in relation to automatic route directions generation.

This chapter is based on: Baltaretu, A., Krahmer, E., & Maes, A. (2016). Landmarks on the move. Producing and understanding references to moving landmarks. *Spatial Cognition and Computation, na.* doi: 10.1080/13875868.2016.1212863

5.1 Introduction

In the last decade, pedestrian navigation systems have become increasingly popular. Augmented reality too, is slowly entering our everyday live. State-of-the-art technology can redefine the capabilities of navigation systems. For example, different devices (e.g., Tesla's car that drives itself safely in a variety of conditions) can capture visual surroundings (with video cameras) in real time. In the future, this could enable pedestrian navigation systems to ground route directions into the visual context and generate instructions by referring to both stable database information (e.g., streets, buildings) and other type of information present in the visual field of the user. One type of information available only in the here-and-now context are moving entities (e.g., cyclists, pedestrians). We know little about how dynamic aspects of the environment can influence the production of route directions in general, and the selection of landmarks in particular. Intuitively, it is likely that, in a co-presence situation, humans help each other with instructions such as "go left where that man turns now". Previous research, however, hardly addressed the issue of moving landmarks, and it is still unclear if and how speakers refer to moving entities and to what extent listeners appreciate such route directions.

By route directions we refer to a set of instructions on how to (incrementally) follow a route (Richter & Klippel, 2004). In this study, we focus on references to landmarks, considered key ingredients for good route directions (Allen, 2000). Traditionally, landmarks are defined as environmental features that function as points of reference (Allen, 2000); "unique configurations of perceptual events [...] (that) identify a specific geographic location" (Siegel & White, 1975, p. 23) or as objects that are better known and that define the location of other points (Presson & Montello, 1988). In general, in previous route direction studies, landmarks are described as concrete, route-relevant, stable entities, such as buildings.

Different scholars theorized that good reference objects are large, geometrically complex and stable (e.g., Campbell, 1993; Talmy, 1983). One likely reason for route direction studies to look upon landmarks as stable entities could be that the communicative situation typically used in the experimental setups includes some type of delay or asymmetry between producing directions and navigating with them. For example, instructions are communicated over distance (e.g., telephone) or asynchronously (on the basis of maps or previous experiences, a participant produces instructions to be later used / evaluated by another one). In such situations, references to here-and-now events are unlikely to be produced. In this study, we focus on a situation in which this delay is absent: turn-by-turn route directions with the instructor and navigator being co-present. While experiencing a shared dynamic environment, speakers can improve the instructions by referring to anything they see.

Arguably, the most important characteristic of a landmark is its distinctiveness. Objects can be distinctive on different dimensions (for example due to familiarity or functional relevance). Here, we focus on navigation contexts that are unfamiliar

to the traveller and in which perceptual salience is more important than other (e.g., knowledge-based) information. It has been theorized that the more visually noticeable or attention-grabbing an object is, relative to neighbouring entities, the more likely it is to be used as a landmark (Sorrows & Hirtle, 1999). For example, colour and size seem to influence landmark selection (Allen, Siegel, & Rosinski, 1978; Sorrows & Hirtle, 1999) as is confirmed by previous experimental work in natural environments (Nothegger et al., 2004; Raubal & Winter, 2002). However, in early processing stages of attention, other visual attributes come into play, such as the direction and velocity of motion (Mital et al., 2011; Treisman & Gelade, 1980). We ask to what extent this attention-grabbing property makes it likely that people refer to moving entities and if instructions with moving landmarks are preferred as much as instructions with more stable 'traditional' landmarks.

5.1.1 Stable vs. moving objects

In the navigation literature, it is common to think of landmark objects as being stable / permanent entities. This stability, as assumed in route direction studies, has beneficial effects on navigation and reorientation. The perceived stability of objects seems to influence toddler's and rodent's use of landmarks for orientation. Studies on toddler's reorientation skills speculate that stability and scale are important factors in landmark use, where smaller or more portable objects have less navigational significance (Learmonth, Newcombe, & Huttenlocher, 2001; Smith et al., 2008). Similarly, rats can search for a location defined by visual landmarks, but will not do so if the landmark's position has varied from trial-to-trial (for a review, see Burgess, Spiers, & Paleologou, 2004). Convergent fMRI evidence suggests that stable objects elicit greater activity in regions of the brain involved in navigation and landmark assignment (for a review, see E. Chan, Baumann, Bellgrove, & Mattingley, 2012).

Most of the experimental route direction studies start with the default assumption that objects have to be stable in order to be used as landmarks and with the exception of the rodent experiments, the studies mentioned above were not designed to test this assumption. In none of them is the object's motion directly witnessed by the participants, and motion is never used as a clue that could potentially help with solving the task. Despite work on direction giving in the context of human dialogue (e.g., Brennan, Schuhmann, & Batres, 2013), as well as detailed analyses of the linguistic structure of route directions (e.g., Allen, 2000), most studies do not address situations in which both the speaker and listener have access simultaneously and can use any aspect of the visual environment.

In typical settings, speakers are asked first to learn routes from text (e.g., Ferguson & Hegarty, 1994; Lee & Tversky, 2005; Tom & Tversky, 2012), from maps (e.g., Lee, Tappe, & Klippel, 2002), by travelling the route (e.g., Denis et al., 2014; Lovelace et al., 1999; May et al., 2003; Miller & Carlson, 2011), or by free recall from memory (e.g., Denis et al., 1999). Then, speakers have to provide a description

and draw sketches for a fictitious listener or take a recognition test. In none of these studies do the means for acquiring and disseminating spatial knowledge allow moving entities to be mentioned. To our knowledge, studies that focus on spontaneously elicited route directions do not make any observations regarding temporarily available information (Couclelis, 1996; Denis et al., 1999; Golding, Graesser, & Hauselt, 1996; Tversky & Lee, 1998). As a result, landmarks are generally understood as stable three dimensional objects, such as buildings and road furniture, and two dimensional, mostly related to the path to be followed (e.g., streets) (Denis et al., 1999; Ishikawa & Nakamura, 2012; May et al., 2003).

Additionally, in the field of automatic landmark selection, buildings are considered prototypical landmarks (Sadeghian & Kantardzic, 2008), and permanence / stability is one of the attributes considered to contribute to the perceptual salience of an object (Duckham, Winter, & Robinson, 2010; Raubal & Winter, 2002). In sum, landmarks are by default assumed to be stable, and different studies (such as the ones on orientation, route direction production and landmark selection) focus on this type of object for practical and methodological reasons.

However, it is likely that moving landmarks are part of the rich repertoire of landmark options, in particular in dynamic navigation situations in which producer and addressee are co-present. Moving entities have various features that make them potentially suitable landmarks. Motion is processed effortlessly by the visual system, and it can efficiently grab and guide attention (e.g., Abrams & Christ, 2003; Hillstrom & Yantis, 1994; Mital et al., 2011). In other words, moving entities are notably perceptually salient (Itti, 2005), a prerequisite for landmarks. In fact, motion contributes to an objects' perceptual salience as much as the combination of all other visual features (colour, size, etc.) (Carmi & Itti, 2006; Itti, 2005).

Among different types of objects in motion, animate entities seem to capture attention even more. By animate entities we refer to entities conceptualized as living beings (Fraurud, 1996). Animates are conceptually highly accessible (e.g., Prat-Sala & Branigan, 2000), and visual representations of the face and the human body have the ability to capture the focus of attention even when visual attention is occupied by other tasks (for a review, see Downing et al., 2004). Humans prioritize the visual processing of animate over inanimate entities (Kirchner & Thorpe, 2006; New et al., 2007), and attention and animacy are linked and bias reference production (Coco & Keller, 2015).

Previous experiences may, however, hinder the use of moving landmarks, and perceptual salience might not suffice to elicit references to moving objects. People are more used to receiving and giving instructions with stable objects. After all, in everyday life, navigation systems currently do not make use of dynamic features and referring to a person that is just turning is an event that requires good timing with the direction giving process. Moreover, task demands can have a significant influence on what people watch and mention. For example, Miller and Carlson (2011) manipulated the objects' perceptual salience (size and colour) and their relevance for the naviga-

tion task (objects placed at decision versus non-decision points). Perceptual salience positively affected object memory, yet it was only the task relevance dimension that determined whether objects were included in the route directions. Congruent evidence comes from Einhäuser, Rutishauser, and Koch (2008) who showed that participants were able to suppress paying attention to moving items when searching for a different object in the scene. These results suggests that motion should be more than attention grabbing. It must be task relevant for both producer and listener. Could motion ease the instruction giving process when showing the correct path?

5.1.2 The current study

In route directions, people nearly always refer to landmarks (Tom & Denis, 2003), and these references maximize the helpfulness of the instructions (Allen, 2000). The type of motion perceived by speakers might influence their referential behaviour. To our knowledge there are no studies that have manipulated motion as a crucial variable in their design. Do speakers refer to moving entities, and more so when they are relevant to the route task? Do listeners have problems understanding and following such instructions? To what extend do people like instructions with moving landmarks compared to stable landmarks and route directions without landmarks?

We propose investigating these questions by manipulating moving entities in a controlled, yet naturalistic environment. Participants are asked to give (Experiment 1) and choose (Experiment 2) route directions informed by dynamic scenes displaying natural every-day events. In an outdoor environment there can be different types of movement (self-produced / induced) of different entities. In this study, we focus on self-produced, animated motion.

In Experiment 1, the speaker and the listener carry out a joint direction giving task, in which the perspectives of speaker and hearer are aligned. This contrasts with previous experimental studies that typically use an asynchronous communication setting, which imposes several spatial and temporal constraints. Based on a video, the speaker is asked to produce a route direction to a co-present addressee. Speakers can refer to any aspect of the visual environment they wish. Apart from moving entities, at each intersection, there are different stable objects available as well.

Just like 'traditional' landmarks, the moving entity should be placed near the location of the navigation action. In this study, there are two experimental conditions depicting moving entities that might differ in relevance for the navigation task: persons moving up to a point where they reach the intersection or taking a turn in the direction in which the listener should also turn. We do not claim that dynamic landmarks would replace the stable ones. Both moving and stable objects might be mentioned in the same utterance (e.g., "turn right at the shop, where that person is turning"). In such cases, the moving entity would help disambiguate the stable landmark with respect to a series of possible distractors (e.g., other shops in the scene).

In Experiment 2, we look at preferences for directions with different types of landmarks. Participants are tested individually in a non-interactive context. They have to choose the route direction they like best out of three options (without landmarks, with stable or with moving landmarks). In line with previous studies that emphasize the importance of landmarks in route directions, we expect participants to choose more often instructions with landmarks over instructions without landmarks. A research question is whether the type of movement would affect preference for the two types of landmarks.

5.2 Experiment 1 - Production

5.2.1 Methods

Participants

112 native Dutch-speaking students from Tilburg University (50 women, 21.2 mean age) participated in exchange for partial course credits. They were paired in 56 dyads (i.e. groups). Participants were randomly assigned to a speaker (35 women) or a listener role. All participants gave written consent for the use of their data.

Materials

The materials consisted of 144 street view HD videos (108 experimental videos and 36 filler videos) recorded in Rotterdam downtown. The experimental videos depicted 36 low traffic, **"+"**- shaped intersections. These intersections have a simple geometric shape, in which just saying "go left", without adding any landmark, would discriminate the target street from the other branches of the intersection. Each intersection was recorded three times illustrating a different motion manipulation as illustrated in Figure 1, corresponding to the three conditions (36 experimental videos per condition): (a) no entities moving towards / coming from the intersection (no movement condition); (b) a moving entity walking / cycling towards the intersection (irrelevant movement condition); (c) the same entity taking a turn in the direction required by the navigation task (relevant movement condition). Note that, since all moving entities may be to some extent relevant due to their proximity to the intersection, the terms 'irrelevant' and 'relevant' motion are used for labelling purposes only. The people recorded were casually walking / cycling down the street, without paying attention to the camera. These people were different from one intersection to another. Thirty-six filler videos were included, capturing a different set of intersections from very crowded pedestrianized areas, where individuals turning were masked by the crowd, as well as intersections with complex geometric structures in which passers-by did not turn in a direction relevant to the navigation task. These fillers were added

to present participants with different navigation scenarios and prevent participants from relying on fixed strategies.

Procedure

Participants were presented with instructions stating that the researchers were developing software that can generate real time/live pedestrian route descriptions based on the visual input coming from a Google Glass video camera and realized in audio format via a smartphone, and that at this stage they were collecting good route directions to further develop their system. Participants were given two paper booklets with line drawing maps of the intersection's shape (the speaker booklet included an arrow showing the direction to be taken at each intersection). The task for the speaker was to provide route directions, while the listener had to mark in his booklet the indicated street. The speaker had to first look at the map, then play the video projection and start giving instructions as soon as possible, while watching the video. The listener had to watch the video at the same time and afterwards mark the intended street on the paper map. The listener was allowed to ask questions if the instructions were unclear. Pointing was discouraged by installing a screen between participants up to shoulder level (see Figure 2). The videos were projected on a white wall, at size of approximately 170×120 cm. Each video lasted about 3 seconds. The videos could not be replayed, but the last frame was displayed until the listener was finished. The video clips were divided across three lists, so that each intersection was shown only once to each participant. Participants were randomly assigned to one of the three presentation lists. The task started with 2 warm-up trials, next 72 trials (36 experimental trials) were presented in randomized order. There were no time constraints.

Design and statistical analysis

This study had Motion Type (3 levels: no motion, irrelevant motion, relevant motion) as within participants factor and Presentation List (3 levels) as between participants factor. For the first analysis, we checked if participants mentioned moving entities and if they did so more often when the motion was task-relevant. The dependent variable was coded as a binary variable: the moving entity was mentioned or not by the speaker in his or her first instruction. Next, we checked if moving entities were mentioned together with a stable object, and we analysed listeners' clarification questions as well as their error rates.

Statistical analyses were performed using logit mixed model analysis (Jaeger, 2008), following the recommendations of (Barr et al., 2013). We used the mixed logit model analysis as it can correctly account for random subject and item effects in one analysis. The models were fitted using the LMER function from the LanguageR

Figure 5.1: Example of experimental clips with no movement (above), irrelevant movement (centre) and relevant movement (below) in Experiment 1

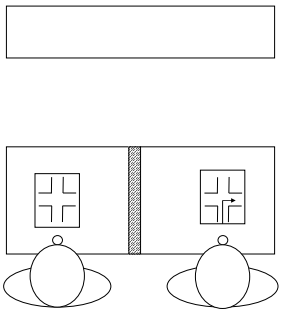


Figure 5.2: Experimental setup in Experiment 1

Package in R (version 2.15.2; CRAN project; The R Foundation for Statistical Computing, 2012). Motion Type and Presentation List were introduced as fixed factors; speakers and video items as random factors. The factors were centered to avoid collinearity. To determine whether the two conditions significantly differed from each other, we started by constructing a model with a full random effect structure. This maximal model included participant and video items intercepts and random slopes to account for between-subject and between-item variation. In case the model did not converge, we only excluded random slopes with the lowest variance until convergence was reached. The results from the first converging model are reported. This model included a random intercept for participants, and random intercept and random slope for Motion Type in item videos. Just as in logistic regression, the beta coefficient is a measure of how strongly each predictor variable influences the dependent variable. The p – values were estimated via parametric bootstrapping over 100 iterations.

5.2.2 Results and Discussion

In total, 2016 route directions (56 speakers $\times$ 36 videos) were produced. Across the three conditions, participants mentioned landmarks ($N = 1113$) in approximately half of the instructions of each condition ($M = 0.48$ in no movement; $M = 0.53$ in irrelevant movement; $M = 0.67$ in relevant movement condition). These landmarks consisted of references to stable objects ($N = 752$), moving entities ($N = 361$), and cases of stable and moving landmarks mentioned together in the same instruction ($N = 43$) (see Figure 3).

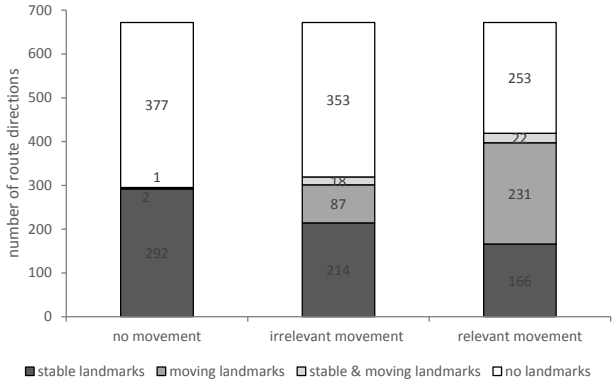


Figure 5.3: Number of route directions with different types of landmarks across conditions in Experiment 1

Moving landmarks

As expected, in the no movement condition, participants rarely referred to moving entities. The three cases in which a moving entity was mentioned in this condition were to a person in the far distance cycling on the road that lead to the intersection. Because of this, statistical analysis was performed only on data from the other two conditions, and revealed a main effect of Motion Type ($\beta = 1.913$; $SE = .27$; $p < .001$). In the relevant movement condition participants referred more often to the moving person taking a turn ($M = .37$, $SD = 0.48$), than in the irrelevant movement condition ($M = .13$, $SD = 0.34$). There was no significant effect of Presentation List ($p > .05$) and no interaction between the two factors ($p > .05$). Some examples of instructions with moving landmarks are given below (English translations of Dutch originals). Speakers focused either on the location of the moving entity (examples 2 and 4) or on the action that the man was doing (example 3). Referring to the pedestrian's location seemed to be the most frequently used strategy in this corpus.

1. You need to turn left, on the same side where the man is walking.
2. Go right, where the man is going.
3. Go on the first street right, follow the cyclist.
4. You need to go left, where the pedestrian is.

Moving and stable landmark combinations

We analysed if moving landmarks were mentioned in the same instruction together with stable ones. In $N = 361$ route directions in which the moving person was mentioned, 11% cases included both moving and stable landmarks ($N = 1$ in the no movement condition, $N = 20$ in the irrelevant movement condition, $N = 22$ in the relevant movement condition, see also Figure 3). The order in which the mixed landmarks were introduced in the instruction was mostly consistent, with moving landmarks being mentioned before the stable ones 80 percent of the times. Below are some examples of mixed landmarks instructions from the irrelevant movement condition (examples 1 and 2) and the relevant movement condition (examples 3 and 4) for which stable and moving landmarks are highlighted:

1. Where *the cyclist* is, go first [street] right, after *the sign with the lion*.
2. Can you see where *the girl* is walking? You need to go between *KFC and Febo*.
3. Go right, first on the right at *the man*, before *the car*.
4. Go straight until *the stairs* and then turn left where *the people* are also walking.

Clarification questions and error rates

Listeners were allowed to ask clarification questions. Their questions could indicate whether or not instructions with moving landmarks were harder to understand and follow.

Sometimes, listeners repeated parts of the route directions while drawing, uttered confirmatory remarks or told the speaker to continue with the next trial. In general, the task was easy, there were few questions and there were no signals of major communication breakdowns in which the speaker and the listener failed to understand each other. Listeners' questions were analysed using the Stivers and Enfield (2010)'s coding scheme for question-response sequences in conversation. According to this scheme an utterance was considered a question if it was a formal question (it had lexico-morpho-syntactic or prosodic interrogative marking) or a functional question (it had to elicit information, confirmation or agreement). Apart from 11 utterances that could not be evaluated due to technical issues, there were 81 questions, posed by 30 listeners. These questions were asked when the speaker did not include any landmark in his initial instruction (49%), when the speaker referred to a stable landmark (30%), a moving landmark (19%) or to both stable and moving landmarks in the same instruction (2%). Compared to the other conditions, there were more questions posed in the no movement condition (44% in no movement, 26% in irrelevant movement, 30% in relevant movement condition). The questions were classified on the basis of the semantic structure of the utterance, and next the question's goal (or social action in Stivers and Enfield's terms) was assessed.

Question type		Goal	Examples	Frequency		
				NM	IM	RM
Polar		confirmation	first?	30%	20%	22%
			left?			
			after Action?			
			am I going like the cyclist?			
Alternative		clarification	left or right?	8.6%	3.7%	3.7%
			I turn in front of the roundabout or go around it?			
			do I turn left or I cross?			
Content	where	information	where is the street?	4.9%	1.2%	3.7%
			where is the bike?			
	what	initiation of repair	what did you say?			2.4%
			what?			

Table 5.1: Frequencies, goals and examples of various listener's questions in Experiment 1, split by condition (NM - no movement; IM - irrelevant movement; RM - relevant movement)

Based on the logical semantic structure of the utterance, questions were classified as polar yes / no question (a question that elicits a confirmation / disconfirmation), alternative questions (questions that included a restricted set of alternative answers), or content (mostly questions seeking information, introduced by a 'WH-word', such as what, where, etc.). The frequency with which these types of questions occurred, the goal for which they were used, and examples are given in Table 1.

These questions might represent problems of various levels of difficulty, and one might expect more content related questions if the task would be difficult. Yet, most of the questions (60 cases) were simple polar questions, asking for confirmation. About half of these (33 cases) were asked after the speaker did not include any landmark in his first instruction. Listeners asked for confirmations with respect to three aspects. There were questions related to the direction of the turn (e.g., 'left?', 16 cases), to correctly choosing the street (e.g., 'first [street]?', 16 cases) and questions about the place where to turn (22 cases). When asking about the place to turn, listeners added references to stable (e.g., 'turn after the shop?', 19 cases) and moving landmarks (e.g., 'turn / go after that man?', 3 cases) or mentioned the path (e.g., 'in the first intersection?', 8 cases).

Second, requests for clarification (12 cases) resulted from minor misunderstand-

ings, although the speakers' instructions generally had landmarks (10 out of 13 instructions had landmarks). Listeners asked for direction clarifications (8 out of 12 cases, e.g., 'left or right?'), and in five of these questions stable landmarks were mentioned. The rest of the questions were about the manner of movement through the intersection (e.g.,'am I on the right or on the left side of the street?'). These might be related to the listener's task, namely drawing the route on paper maps, and listeners wanted to know in more detail how to move in intersections.

Third, there were few requests for information. There were six 'WH' questions, one including a moving landmark reference. In sum, listeners posed questions primarily to confirm their choice, and to a lesser extent, to find out more information about the street they should follow.

Lastly, listeners made few errors when choosing the street indicated by speakers. There were only eleven cases in which they marked a wrong street on their maps and another eight cases in which they corrected their initial choice. The small number of errors suggests that the task was simple and listeners had no real problems accomplishing it.

To sum up, these results revealed that speakers mentioned moving entities when giving route directions. They especially did so when the movement trajectory was informative for the place where a turn should be taken. Moving entities were rarely mentioned together with stable ones, which might indicate that moving entities can be seen as an alternative type of landmarks. Listeners had no difficulty to understand and use instructions with moving landmarks. They rarely asked for clarification and made few errors choosing the correct street, after hearing an instruction with a reference to a moving target. In particular, there was a similar number of clarification questions when speakers referred to stable and to moving landmarks, which makes it interesting to look in more detail at participants' preferences for these entities. To do so, in Experiment 2, we showed videos depicting irrelevant and relevant movement and asked participants to choose the instruction they prefer most.

5.3 Experiment 2 - Evaluation

5.3.1 Methods

Participants

Thirty two native Dutch-speaking students of Tilburg University (12 women, 20.7 mean age) participated in exchange for partial course credits. None of them participated in Experiment 1. All participants gave written consent to the use of their data.

Materials

The materials consisted of 72 videos (the experimental trials from the irrelevant and relevant movement conditions) used in Experiment 1. Overlaid on the videos, a semitransparent red arrow depicted the route and the direction to be followed (see Figure 4).

For each video, we created a set of three route directions as follows: First, we analysed the data coming from ten speakers from Experiment 1 that referred most often to landmarks (irrespective of condition and type of landmark mentioned). The directions could have different information structure (see Experiment 1 for examples). The most frequent word order they used (about 80% of the cases) consisted of a verb and the direction of turn, followed by a landmark. Next, starting from this dominant template, we created for each video three different route directions: one without landmarks, consisting only of a motion verb and the direction (e.g., "turn left"); a route direction with a stable landmark (e.g., "turn left at the Hema shop") and a route direction with a moving landmark (e.g., "turn left where that man / woman / cyclist is going"). In Experiment 2, all the instructions had the same information structure as the one most frequently observed in Experiment 1, namely a verb + direction + stable (e.g, "at the pharmacy") or moving landmarks (e.g., "where the [man / woman / cyclist] is going"), in this exact order. The stable landmarks used in these route directions were the most often mentioned stable objects in Experiment 1. For the moving landmarks, we used the most frequent referring expressions (the man / the woman / the cyclist).

Procedure

As in Experiment 1, participants were presented with instructions stating that the researchers were developing software that can generate real time/live pedestrian route descriptions based on the visual input coming from a Google Glass video camera and realized in audio format via a smartphone. The participants' task was to evaluate different route directions produced by this application. Participants had to watch a number of videos, and for each read the three route directions and choose the one that they liked most. The videos could not be replayed, but the last frame of the video was visible until the participants clicked on the route direction of their choice. Two presentation lists were created, so that each intersection was shown only once to each participant. The trials consisted of a fixation cross displayed for 500 ms, followed by a slide that displayed the video on the upper three quarters of the screen, and three instructions placed below the video, next to each other. The position on screen of the three types of route directions was counterbalanced and randomized, so that each type would be displayed an equal amount of times on the left, centre and right side of the screen. Once the choice was recorded a new trial would automatically start. The experiment began with two practice trials, followed by 36 experimental trials presented in random order. There were no time constraints.

Figure 5.4: Example of experimental clips with irrelevant movement (above) and relevant movement (below) with arrows showing the turning direction in Experiment 2

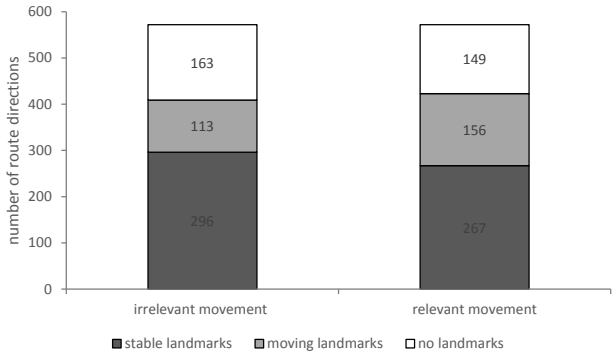


Figure 5.5: Number of route directions types chosen across conditions in Experiment 2

Design and statistical analysis

This study had Motion Type (2 levels: irrelevant motion, relevant motion) as within participants factor and Presentation List (2 levels) as between participants factor. The dependent variable was the type of route direction chosen. Statistical analysis was performed as in Experiment 1. Motion Type and Presentation List were included as fixed factors; subjects and videos as random factors; random intercepts and random slopes for speakers and videos. The first converging model is reported (p – values were estimated via parametric bootstrapping over 100 iterations). This model included random intercept for subjects and random intercept for videos.

5.3.2 Results and Discussion

Out of 1152 cases (36 scenes $\times$ 32 participants), route directions with landmarks were chosen more often (73% of the cases) than route directions without landmarks (see Figure 5). In order to check if motion influenced the choice for a specific type of landmark, the statistical analysis was done on a data set consisting of only the route directions with landmarks.

There was a main effect of Motion Type ($\beta = 1.211$; $SE = .265$; $p < .001$). For videos depicting irrelevant movement, participants chose more often instructions with stable landmarks ($M = .85$) than with moving landmarks ($M = .15$). For

videos depicting relevant movement, the same pattern is observed (stable landmarks $M = .75$; moving landmarks $M = .25$). There was no significant effect of Presentation List ($p > .05$), and no interaction between the two factors ($p > .05$).

There was a slight increase in the preference for moving landmarks in the relevant movement condition. In order to test if this observation is statistically significant we run a Wilcoxon matched-pairs test. There was a significant difference in the scores for irrelevant movement ($M = .15$) and relevant movement ($M = .25$), $z = -2.67$, $p < .05$, $r = -.65$.

In conclusion, participants preferred instructions with landmarks over instructions without landmarks. Stable landmarks were chosen more often than moving landmarks, as the preferred ones. Compared to the ease with which listeners used moving landmarks in Experiment 1, the overwhelming preference for stable landmarks is striking.

5.4 General Discussion

In this study, we manipulated the trajectory of moving entities walking in intersections and analysed if and when speakers refer to moving entities (Experiment 1) and participants' preferences for these references (Experiment 2). In the first experiment, we used an interactive setting, in which speakers were giving instructions to listeners, the two having the same visual information and access to the same navigation environment (videos of intersections). Speakers included (both types of) landmarks in approximately half of the instructions. Human speakers do refer to moving landmarks in the co-presence situation. Motion is perceptually salient, however not only salience, but also task relevance greatly contributed to our results (see also Miller & Carlson, 2011). Moving landmarks were mentioned especially in the relevant motion condition. In fact, in this condition, there were more references to moving than to stable landmarks. Quite often the moving entities were mentioned alone without further references to stable objects, suggesting that moving items can be used as landmarks on their own. Listeners did not encounter problems understanding such instructions, as revealed by the analysis of their questions and error rates. They (rarely) asked for clarifications, and the questions they did ask were mostly due to minor misunderstandings.

In the second experiment individual participants had to choose an instruction for videos depicting irrelevant and relevant motion. Instructions with landmarks were preferred over instructions without landmarks. This result is in line with previous studies that underline the importance of landmarks at decision points where reorientation is required (Denis et al., 1999). In addition, route directions with stable landmarks were preferred over instructions with moving landmarks. Below, these results are further discussed in relation to route direction production, and possible suggestions for automatic route direction generation are proposed.

5.4.1 Moving landmarks in route directions

In earlier research, landmarks are considered stable entities, and there are indeed many communicative situations where stable objects are valid, relevant landmarks (route directions from maps, memory, etc.). The asynchronous situations used in previous studies favour an approach that defines landmarks as objects that are stable, memorable and stand out in the general knowledge one has about the environment. Yet, moving entities can be useful landmarks under specific circumstances (direction giving as a joint activity and while observing moving entities during utterance planning). They seem to be effective in route directions due to their perceptual salience, but more so when informative. In our setup, moving entities are relevant because they show the right turn in the intersection. Since motion involves both spatial and temporal dimensions, the main role of these moving entities was to provide short-term orientation for the listener and locate with their presence the street where a turn should be made. In addition, the joint perception of an entity moving on a relevant trajectory affords the visualization of an emerging action: the future turning action of the listener. An instruction such as 'follow the man in red' can subsume several pieces of information. It tells the listener the place where to turn, and the direction in which the turn should be done, and it singles out the street on which to continue.

In order to evaluate how listeners cope with such instructions, we analysed the listeners' clarification questions and their error rates. In general, there were few errors and the task was simple. Listeners rarely asked clarification questions and their drawings were mostly correct. Interestingly, most of the questions were asked when the speaker's instruction did not contain landmarks. Listeners introduced both moving and stable landmarks in their questions, however stable landmarks predominated. There could be different explanations for this choice, such as a possible listener's preference for stable landmarks or an effect of the timing of the events: by the time listeners formulated their questions, the entity would have moved out of sight.

To further test if moving landmarks were considered as good as stable ones, we asked a different group of participants to watch videos depicting relevant or irrelevant motion and choose the instructions they liked best. They were presented with three instructions: without landmarks, with stable or with moving landmarks. Participants preferred instructions with landmarks over instructions without them. Stable landmarks were chosen more often than moving landmarks. When motion was task-relevant, the preference for moving landmarks slightly increased. Yet, the increase in preference for moving landmarks was rather small. Given that listeners can adapt and use moving landmarks fairly easily, the preference for stable landmarks might have been influenced by the non-interactive context in which the task took place. Another possible explanation could be that there is a speaker - listener asymmetry. Movement might primarily capture speaker's attention, and speakers

do not necessarily tailor their utterance to the listener's needs. Further research could evaluate a possible effect of the joint communicative situation on landmark preference.

Overall, the data collected in the two experiments suggest that landmark references were influenced by dynamic events present in the environment. However, one could expect that the frequency with which moving landmarks are mentioned depends heavily on the scene. The scenes studied in these experiments are admittedly relatively simple. For example, there was only one moving entity present in each scene. Real environments are noisier, and we can assume that traffic and visual clutter (e.g., surrounding pedestrians and cars) would influence the choice of landmark objects. An interesting question for future research is how more complexity (e.g., multiple moving entities, complex intersections) affects the usability of moving landmarks.

If an entity moves in the relevant direction, the speaker would have to minimally say "follow it", avoiding the burden of describing the turn or the street. This might ease the speaker's task, especially when describing complex turning actions (Klippel et al., 2013). For the addressee, following another person may be beneficial, as well (Krieg-Brückner, 1998, for 'following' as a basic navigation strategy). Moving landmarks bear some similarity to linear landmarks (Richter & Klippel, 2004), which afford being followed (e.g., "follow the river"). Yet, a moving entity can only guide in one specific direction (unlike linear landmarks, such as rivers, which stretch in two directions), and they are not suitable for determining actions in several decision points (e.g., "follow that person for three intersections and then go left").

Any navigation help used in route directions might be considered a landmark (Couclelis, Golledge, Gale, & Tobler, 1987; Richter & Winter, 2014). Landmarks could range from personal events, as long as the latter are salient in memory and part of our shared knowledge (e.g., I tripped on the kerb, and I broke my ankle, and you know this place), to concrete, stable objects that have been discussed in the literature. In between, there could be other types of landmarks, such as intersections, linear (rivers) and area like (forests) objects. A moving landmark might be placed somewhere between these classes: it can attract attention like a classical landmark, but it is in some sense personal, because it is an ephemeral event witnessed by a small group of people. Experimentally study of these aspects is much needed before having a more clear position regarding the role of moving entities in broader theories on navigation.

Finally, it is worth investigating if the effect observed is due only to movement being relevant or also to animacy. Would speakers still mention the moving landmark if this would be a car? We suspect it is movement and showing the right turning direction that matters.

5.4.2 Implication for route directions generation

The automatic generation of route descriptions is often studied within Natural Language Generation, NLG (e.g., Dale et al., 2005; Roth & Frank, 2009). NLG systems typically involve Referring Expression Generation (REG) (e.g., Krahmer & van Deemter, 2012) for generating references to landmarks. Until recently, REG for landmarks and studies of route direction production have focussed exclusively on references to stable entities. Given the results of this study we will sketch some suggestions regarding automatic generation of moving landmark references.

Though there is room for improvement, human detection and tracking in videos has reached a quality level that would allow navigation systems to make use of dynamic aspects from the environment (Dollar, Wojek, Schiele, & Perona, 2012; Venugopalan et al., 2015). For generating more human-like instructions, navigation systems could include references to moving objects in route directions, as long as estimations of the eye gaze are available to track what the listener is watching and the instructions are well timed with the event. Our data shows that speakers spontaneously refer to items with a task relevant trajectory. Incorporating references to dynamic landmarks could optimize route directions, by providing a wider range of possibilities for choosing landmarks. For example, buildings might not be present or are hard to refer to. Sometimes buildings do not present any particularity (such as shops or architectural features) and might have indefinite colours or a small size contrast with other buildings. In such situations, moving landmarks seem a natural option and listeners should successfully handle such instructions.

Two aspects pose further challenges for integrating this type of reference in route directions. First, our data showed that in a non interactive context participants preferred stable landmarks. It is an empirical question if users engaged in dialogue with a system would consider moving landmarks as good as stable ones. The efficiency of these instructions during navigation needs to be assessed. Second, little is known about the relation between motion and reference processes, and this poses specific challenges for REG.

Typically, REG algorithms produce references by using a predefined list of preferred attributes. An object is first referred using a noun (e.g., go left at the building). If this description does not single out the object from the scene (there are several buildings), more attributes (e.g., colour, size) are added until all the other objects are eliminated. Some of these attributes are used more often (are preferred) over others (e.g., colour is often mentioned, unlike size or location) and algorithms would typically make use of this preference (Dos Santos Silva & Paraboni, 2015; Krahmer & van Deemter, 2012). However, motion has been shown to attract visual attention even more than colour, intensity or orientation features (Carmi & Itti, 2006; Itti, 2005; Mital et al., 2011). More experimental work is needed to investigate how motion relates to these visual cues. Many studies that examined static scenes provided valuable insights on how perceptual saliency affects attribute preference (e.g., Clarke,

Elsner, & Rohde, 2013; Viethen & Dale, 2008), yet the scalability of their conclusions to the dynamic world remains an open question.

Finally, cutting edge technology provides machines with a rich sensory input. Our results suggest that movement detected in the nearby environment can be informative in a landmark selection task. These results highlight that not only the stable landmarks, but also the moving entities play a role in the production of turn-by-turn route directions.

CHAPTER 6

Conclusions and Discussion

In this dissertation, we reported on four studies in which we investigated speakers' referential choices in spatial domains, with a focus on references to landmarks, using various kinds of visual scenes, including photorealistic street images and video's. We studied how a range of factors influenced the production and comprehension of spatial references, including visual properties such as perceptual salience (Chapter 2 and 5) and visual clutter (Chapter 4) and the communicative task, including the purpose of the interaction (Chapter 3) and the complexity of the task (Chapter 4). Even though the results have been discussed at length in the corresponding chapters, in this concluding chapter, we give a summary of all four studies and discuss general theoretical implications of this work. We end with suggestions for future studies and a general conclusion.

6.1 Visual properties, a summary of the empirical findings

Throughout the studies presented in this dissertation, our results suggest that visual properties influence referential processes in spatial domains. We have analysed visual properties of objects-to-be-described (e.g., perceptual salience) and visual contextual variables (e.g., the extent to which a navigation scene has a high level of visual clutter). Our findings can be summarized as follows.

Firstly, in Chapter 2, we found that the spatial position of a relatum object systematically influenced relatum reference, while salience only had a minor impact on reference production and comprehension. In this study, speakers were asked to refer to a target object, and in order to do so unambiguously, they had to describe the object in relation to one or two relatum candidate objects. Our results showed that leftmost positioned entities were referred as (first) relatum relatively more often, and this result was consistent across four production experiments. Leftmost entities were chosen in approximately 60 percent of the stimuli (with 50 percent representing a random selection). This moderate frequency suggests that other factors, which in general are known to influence language production, could also influence relatum reference. In line with previous studies (Viethen & Dale, 2008; Kelleher et al., 2005; Vogels et al., 2013), we expected that the objects' (conceptual and perceptual) salience would influence relatum reference. Surprisingly, this factor yielded no effects. From a listener's perspective, the formulation of the description and the position of the animate entity in the scene did somehow influence the acceptability ratings given to descriptions. This contrasts with the production data, suggesting that what speakers do is not necessarily what addressees appreciate.

We discussed, in Chapter 2, various possible reasons for the lack of an effect of animacy and visual salience on reference production. The results were consistent with Viethen and Dale (2008) and Viethen et al. (2011), who also had reported limited

effects of relatum salience, in scenes containing a small number of objects such as cubes and spheres. Yet, there is also growing evidence that perceptual salience does influence reference production and comprehension, especially in visually complex scenes, where speakers prefer to talk about objects that attract attention (Clarke, Elsner, & Rohde, 2013; Clarke et al., 2015). It might be the case that, with more natural stimuli, the factors investigated in Chapter 2, could have shown stronger effects. Hence, in the later chapters of this thesis, we have investigated the influence of visual properties on reference production in more realistic settings, which better approximate the level of visual detail of every-day life situations, and within the specific context of a goal directed task: route directions.

An important characteristic of naturalistic scenes is the frequent presence of visual clutter. Visual clutter can be used as a proxy for estimating the number of objects in a scene (Rosenholtz et al., 2007), and it is known to affect vision processes, such as object recognition performance (Bravo & Farid, 2006), scene segmentation (Bravo & Farid, 2004), and visual search (Henderson et al., 2009), as well as language and reference production (Coco & Keller, 2009; Koolen et al., 2013; Clarke, Elsner, & Rohde, 2013). In our studies, we expected that a high level of visual clutter would cause more uncertainty regarding the navigation task, and thus speakers would give more detailed instructions (e.g., containing a higher number of landmark references, which could help disambiguate the street that needs to be taken) and this type of instructions would be beneficial and preferred by the addresses. Indeed, our results showed that visual clutter can substantially affect reference production and addressees referential preferences (Chapter 4). We found that high levels of visual clutter triggered more references to landmarks and that addressees were more likely to prefer such instructions for the complex visual scenes. We did not find an effect of visual clutter on the speed with which the addressee selected the street on which to continue, which might have been caused by the way we graphically marked these streets in the visual scenes. However, it could be the case that visual clutter affects to a larger extent speakers, rather than addressees. Depending on how speakers formulate their references (e.g., by mentioning the most perceptually salient things first), addressees might have an advantage in finding the target (Clarke et al., 2015).

Of particular interest was the fact that some of the objects chosen as landmarks were atypical, such as pedestrians and moving or parked cars. According to much of the previous literature on the topic, landmarks are, almost by default, stable entities, yet we observed that motion is becoming more relevant in online in situ navigation, where a lot of movement can be observed. In general, we know that movement attracts attention, which in turn could contribute to the perceptual salience of an object. In Chapter 5, we therefore studied references to moving entities in more detail. In the first, route direction giving experiment, speakers were found to include in their instructions both references to stable landmark objects (buildings), as well as to moving entities (pedestrians walking in the proximity of the intersection or taking a turn). Moving entities were mentioned often and without further references

to stable objects. This result suggests that moving entities could be considered as landmarks on their own. Further work is necessary for testing the extent to which speakers make use of these entities in 'the wild' (see Hölscher, Tenbrink, & Wiener, 2011). In a second, evaluation experiment, we expected instructions with landmarks to be preferred over instructions with no landmarks, which was indeed the case; and we observed that route directions with stable landmarks were overall preferred over instructions with moving landmarks. It might be the case that addressees prefer what they are familiar with, directions containing stable entities, or maybe this preference is related to the non-interactive nature of the task.

6.2 Task-related aspects, a summary of the empirical findings

Apart from visual properties, in this thesis we also investigated task-related aspects which might influence the production and comprehension of referring expressions. Often, prior referring expression production experiments have only focussed on identification. As a result, it is uncertain how generalizable their results are to other communicative tasks, going beyond identification. Thus, in Chapter 3, we studied whether referring to the same target could differ depending on the communicative purposes of the interaction (object identification vs. giving route directions). We expected that having different purposes might bring into attention different aspects of the target, and a distinct focus of attention could influence the particular choice of object's properties to be included in a referring expression (Beun & Cremers, 1998). We indeed found noticeable formulation differences, with longer and more detailed referring expressions produced in the identification task than in the route directions giving task. This result suggests that different communicative purposes require different strategies. Yet, the results showed no semantic differences regarding the content of the referring expressions produced in the two tasks. Speakers, in both tasks, used the same types of attributes (e.g., location and colour) almost equally often. This latter finding is further supported by a separate evaluation experiment, which showed that participants had no systematic preferences for a specific type of phrase.

In Chapter 4, we studied how the complexity of the task might influence references in route directions and we found an interaction with the level of visual complexity. We had one main expectation: we predicted that the intersection structure would influence the way speakers disambiguate a target street, by including more path and landmark references. We distinguished between simple intersections (e.g., +- shaped, four branch intersection) and complex ones (e.g., a **K**-shaped intersection or intersections with more than four branches). When a speaker needs to describe a right turn, the latter types were expected to be conceptually more complex due to the turning angle and the number of branches. Indeed, in those cases speakers

produced more path and more landmark references, although this was qualified by an interaction with visual clutter; if the environment was more complex, speakers were more inclined to refer to surrounding objects. Instructions with landmark references were found to be beneficial for addressees that had to choose a street in a complex intersection.

6.3 Implications for automatic route directions generation

The results presented above have several implications for automatic generation of references in spatial domains. As explained in the introduction of this dissertation, as well as in various chapters, most automatic route direction generation engines have a module (Referring Expression Generation, REG) responsible for the production of references to the target streets and landmarks. Commercial systems are mostly limited to producing references to street names, which does not necessarily resemble what human speakers do in a navigation scenario (Tom & Denis, 2004). Most algorithms, used for generating route directions, and more generally in the REG field, do not actually take into account the visual context in which references are generated. Rather, these systems tend to operate on databases consisting of semantic annotations, approximating the visual context. These semantic annotations do not capture a broad range of perceptual features and relations among objects. Thus, current REG algorithms account poorly for how people refer in visual spatial domains. More recently there have been some proposals on how to account for the visual context when generating referring expressions (e.g., Mitchell, van Deemter, & Reiter, 2013). Automatic comprehension of referring expressions seems also to be boosted by modelling relationships between objects in the visual context of a scene rather than modelling only target object properties (Nagaraja, Morariu, & Davis, 2016). Taking into account the visual context has implications for when to add references to objects and how to refer to them.

Based on our results, we can formulate three implications for REG. First of all, we suggest that findings regarding content selection from studies where identification was the main purpose of the interaction (van Deemter et al., 2006; Clarke, Elsner, & Rohde, 2013; Koolen et al., 2013), could generalize to a large extent to other settings, such as landmark reference generation, in particular where the selection of attributes is concerned. We base this suggestion on the fact that we did not find semantic differences between references produced for identification purposes and those produced during route direction giving in Chapter 3.

Second, our results suggest that, when a REG algorithm needs to produce a relational description, the structure of the scene becomes a relevant starting point in guiding the choices for relatum reference generation. When the distance between the

target and multiple relatum candidates is comparable, algorithms should take into account the spatial position of the object, as well as the relatum objects' salience. In situations in which there are several relatum candidates similarly aligned, we suggest that entities placed on the left of the target should be favoured (assuming that the system's goal is to generate references in a humanlike fashion). Moreover, we suggest that locative information should be one of the most preferred (and hence most frequently used) property, next to colour, when referring to targets in naturalistic scenes. Our findings (Chapters 3 and 4) contradict the common assumption that locative information is a 'dispreferred' property, to be used by an algorithm as a last resort (Dale & Haddock, 1991; Krahmer & Theune, 2002; Krahmer et al., 2003).

Third, in 'simple' situations (simple intersections, a low cluttered environment), landmark references are not required. Instead, navigation systems should add references to landmarks when the level of visual clutter is high, irrespective of the type of intersection; and more path and landmark references in situations with complex intersections and cluttered environments. A system that makes use of landmarks in complex situations might reduce navigational uncertainty and help users make correct choices. Finally, navigation systems could include references to dynamic landmarks. When stable landmark objects do not present much distinctive particularity (e.g., indefinite colours or a small size contrast with other buildings), moving landmarks seem a natural option and listeners should successfully handle such instructions.

6.4 Future research

We would like to end with some observations regarding central themes that have emerged across different chapters and that, we believe, provide interesting lines for future research. A key element of this thesis is the transition from stimuli and tasks typically used in 'identification studies' towards more visually complex stimuli and a more natural discourse context (giving route directions). These two aspects, we believe, are worth more detailed, further investigations.

First of all, throughout this thesis, we have shown that when referring in a spatial domain, scene complexity and perceptual salience influence referential processes. As discussed above, scene perception is typically not taken into account by REG algorithms. Similarly, most computationally psycholinguistic studies on reference have used artificial visual scenes, consisting of grids of unrelated objects (e.g., TUNA corpus, van Deemter et al., 2006) or simplistic scenes (e.g., GRE7D corpus, Viethen & Dale, 2011). In order to develop REG algorithms that can account for how people refer in natural spatial domains, it would be interesting to know more about the interplay between language and vision at the level of reference production. This raises various questions. For example, to what extent do speakers pay attention to other objects in the scene (as potential distractors) and compare the target to nearby objects? While most algorithms compare the target's properties to the properties of all the

other objects in a scene, it is very unlikely that speakers pay attention to each building and each object when giving directions in a naturalistic, cluttered environment. Yet, some comparisons to nearby objects are likely to happen, as speakers choose to mention relata and other properties that are unique for one building, but not for the neighbouring ones.

In general, a better understanding of how visual processes interact with reference production (and language production in general) is needed: how do speakers choose attributes and relations between (landmark) objects, and the actions in which these objects are involved? We would like to know not only where speakers look during scene viewing, but also what is perceived by the speaker. To what extent could we use models of visual attention to predict the content of a reference (see Clarke et al., 2015), and to redefine the notion of distractor set in complex scenes to a restricted area around the target? To what extent could models of perceptual salience inform models of reference production when the task is carried out in a real dynamic environment, outside of the 'watching-a-computer-scene' paradigm? Answering questions such as these will not only lead to potentially better route descriptions, but can also improve our understanding of language production in visually contexts.

A second theme for future research concerns the nature of the interaction between the navigation system and the user. So far, we have mainly assumed that the navigation system merely produces its descriptions, and the user just listens to them. But more interactive options are conceivable as well, and this is likely to influence reference production. Many studies looking at the computational production of referring expressions have focussed on so-called "one-shot" descriptions. These do not question the basic assumption that the discourse context might influence the way objects are referred to, and most REG algorithms are not truly able to model the discourse context dimension (Krahmer & van Deemter, 2012). Previous studies pointed out that interaction is one factor that influences referring expressions: speakers align and tailor their references for the addressee and repeated references are reduced (e.g., Brennan & Clark, 1996; Pickering & Garrod, 2004). Our results suggest that even basic aspects, such as the communicative purpose affects the referring expressions formulation. Focusing on interactive, situated settings for REG in the navigation domain would make an interesting line for future research. Future navigation systems would not need to rely on one-shot descriptions, but take user feedback into account (either explicitly, by posing clarification questions "which building do you mean?" or implicitly, by tracking the gaze of the user for understanding, cf. Garoufi, Staudte, Koller, & Crocker, 2015). As above, this has the potential of improving navigation systems, but would also further our understanding of reference production in more natural, interactive settings (see also Mast, Smeddinck, Strotseva, & Tenbrink, 2010; Ross, 2011).

6.5 Conclusion

This dissertation has shown that visual properties and task-related aspects affect definite object descriptions in spatial domains. Scene properties, such as visual clutter and properties that contribute the visual salience of objects were shown to affect when speakers refer to objects from the environment (landmarks) while giving route directions, as well as the semantic content they select for their references and the type of objects that speakers choose to refer to. The purpose of the interaction affected the formulation of referring expressions and the complexity of the task influenced the frequency with which landmark references were produced. Based on the empirical findings, we have formulated recommendations for automatic generation of landmark references and suggested two lines for future research related to discourse context and naturalistic visual environments. We hope that our findings can be helpful for future generations of navigation systems, both for car drivers (assuming that humans will continue to drive their own cars) and for pedestrians. And perhaps some day, when you are walking through a city centre and a cyclist captures your attention, your navigation system will tell you "Follow the cyclist!"

Summary

It's probably fair to say that we don't spend much time thinking about the way we talk about the space around us. Yet, we frequently make use of spatial references. For example, when giving route directions, we often add spatial references to entities from the environment, also known as landmarks ("go straight until you see a *pharmacy on the left*"). References to landmarks are an important element of route directions produced by humans. Automatically generated references often lack this feature. By acquiring a better understanding of when and why humans use landmarks, we create conditions for the development of more human-like algorithms for generating route directions. Route directions can be produced off-line (e.g., on the basis of maps) or step-by-step while navigation unfolds in the here-and-now context. They are produced in environments differing in visual complexity and types of landmarks (e.g., a busy city centre with moving cars and pedestrians vs. a quiet residential neighbourhood). Both humans and machines need to take a series of decisions: when to add a reference to landmark, how to refer to it and what type of objects can be considered good landmarks for directions? Up to date, we still have a poor understanding of how speakers produce and understand reference in visually complex spatial domains.

In this thesis, we describe several psycholinguistic experiments in which we study how speakers produce referring expressions and how listeners interpret them. We focus on semantic aspects (what type of information is included) and pragmatic ones (why do speakers include this information). Moreover, we attempt to formulate implications for developing natural language generation algorithms that could automatically produce human-like route directions.

Study 1

In the first study (Chapter 2), we started by analysing the extent to which a set of basic features (spatial position and salience) influence the production of relational descriptions (such as "the ball between the man and the drawer"). In Experiment 1, speakers were asked to refer to a target object (a ball) and in several studies we addressed the role of spatial position, more specifically if speakers mention the entity positioned leftmost in the scene as (first) relatum. The results showed a small, but robust preference to start with the left entity. This suggests that other factors could influence spatial reference. In the following studies, we varied salience systematically, by making one of the relatum candidates animate and by adding attention cues, subliminally and then explicitly. There was no evidence for an effect of visual salience and little evidence that animacy plays a role. In Experiment 2, we tested the acceptability of the referring expressions. Participants expressed their preference for specific relata, by ranking descriptions on the basis of how good they thought the descriptions fitted the scene. Results showed that participants preferred most the description that had an animate entity as the first mentioned relatum. In the next three chapters, we gradually turned to more complex visual scenes and continued testing possible effects of the visual context and task-related aspects on references.

Study 2

In the second study (Chapter 3), we have investigated how the speaker's communicative purposes (giving directions or identifying objects) might influence the content and formulation of references. In Experiment 1, speakers referred to a target building nearby or further away, so that their addressee would distinguish it between other buildings (identification) or give route directions and use the same building as a landmark (instructions). Our results showed that irrespective of the speaker's purposes, referring expressions consisted of the same types of attributes, yet the attribute frequency and formulation differed. In the identification task, the referring expressions were longer, contained more locative and more post-nominal modifiers. In Experiment 2, a different group of participants had to evaluate references produced in Experiment 1, while assessing descriptions of objects or descriptions of objects extracted from route directions. Neither task, distance, nor the length of the phrases influenced their choice, indicating that addressees consider references produced in both conditions equally adequate in both uses.

Study 3

In the third study (Chapter 4) we tested if references to paths and landmarks in route directions could be influenced by environmental complexity. In this study we focused on two aspects of the visual surroundings: namely intersection structure and visual clutter. Speakers were asked to produce (Experiment 1), understand (Experiment 2) and evaluate (Experiment 3) turn-by-turn route directions in a naturalistic setting (Google Street View panoramic pictures). Our results showed that increased levels of visual clutter and intersections with complex structures trigger more references to landmarks and paths when participants produce directions, longer decision times to determine what the next correct step in a route is, and increased preference for landmarks.

Study 4

In the forth study (Chapter 5) we have investigated if and when speakers refer to moving entities in route directions and how listeners evaluate such instructions. While there is a general agreement that landmarks should be perceptually salient and stable objects, other attributes, such as (animated) motion, can also attract visual attention and make entities salient. We asked speakers to watch short videos of different crossroads with and without moving landmarks and give directions to listeners, who in turn had to choose a street on which to continue (Experiment 1) or choose the instruction they most preferred among three route directions (Experiment 2). Results revealed that speakers had mentioned moving entities, especially when the trajectory was informative for the place where a turn should be taken (Experiment 1). Listeners had no problem understanding instructions with moving landmarks (Experiment 1). Yet, participants chose instructions with stable landmarks more often (Experiment 2). These results are discussed in relation to

automatic route directions generation.

Conclusion

In this thesis, we showed how a range of factors influenced the production and comprehension of spatial references. We have tested visual properties such as perceptual salience (Chapter 2 and 5) and visual clutter (Chapter 4) and the communicative task, including the purpose of the interaction (Chapter 3) and the complexity of the task (Chapter 4). All these factors have been shown to affect semantic or pragmatic aspects of spatial reference.

Moreover, we tried to formulate implications for algorithms in the field of Referring Expression Generation. Our studies suggest that (1) results from previous studies that use identification tasks would generalize to other contexts, (2) locative information is probably more important than previously thought and the spatial structure of a scene could be used as a starting point for choosing a referent in a relational description (3) visual complexity is informative in deciding if a landmark is required or not.

AcknowledgeMoments


TILBURG
UNIVERSITY

Publication List

Journal Publications (peer-reviewed)

Baltaretu, A., Krahmer, E., & Maes, A. (2016). Referential choice in identification tasks and route directions. *under review.*

Baltaretu, A., Maes, A., & Krahmer, E. (2016). Landmarks on the move. Producing and understanding references to moving landmarks. *Spatial Cognition and Computation: An Interdisciplinary Journal.* DOI: 10.1080/13875868.2016.1212863.

Baltaretu, A., Krahmer, E., van Wijk, C., & Maes, A. (2016). Talking about relations: Factors influencing the production of relational descriptions. *Frontiers in Psychology*, *7*(103). DOI: 10.3389/fpsyg.2016.00103.

Baltaretu, A., Krahmer, E., & Maes, A. (2015). Improving Route Directions: The Role of Intersection Type and Visual Clutter for Spatial Reference *Applied Cognitive Psychology*, *29*(5), pp. 647–660. DOI: 10.1002/acp.3145.

Conference papers (peer-reviewed)

Baltaretu, A., & Castro Fereirra, T. (2016). Task demands and individual variation in referring expressions. In A. Isard, V. Rieser, D. Gkatzia (Eds.), *Proceedings of the 9th International Natural Language Generation conference.* Stroudsburg, PA: Association for Computational Linguistics.

Baltaretu, A., Krahmer, E., & Maes, A. (2016). Referential choice in identification tasks and route directions. In D. Grodner, D. Mirman, A. Papafragou, J. Trueswell, J. Novick, S. Arunachalam, S. Christie & C. Norris (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Baltaretu, A., Krahmer, E., & Maes, A. (2015). Moving Targets: Human References to Unstable Landmarks. In A. Belz, A. Gatt, F. Portet, & M. Purver (Eds.), *Proceedings of the 15th European Workshop on Natural Language Generation.* Stroudsburg, PA: Association for Computational Linguistics.

Baltaretu, A., Krahmer, E., & Maes, A. (2015). Landmarks in motion: Unstable entities in route directions. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Baltaretu, A., Krahmer, E., & Maes, A. (2014). Following route directions: The role of landmark reference, intersection type and visual clutter. In A. Eshghi, K. Fukumura, & S. Janarthanam (Eds.), *RefNet workshop on psychological and computational models of reference comprehension and production.* Edinburgh: RefNet.

Baltaretu, A., Krahmer, E., & Maes, A. (2014). Designing context aware instructions: Perceptual salience and task demands in the selection of natural landmarks. In E. de Vries (Ed.), *Proceedings of the Bi-Annual Conference of EARLI Special Interest Group Comprehension of Text and Graphics, Bridging representations.* Rotterdam: Erasmus University.

Baltaretu, A., Krahmer, E., & Maes, A. (2014). Route Descriptions: The Role of Intersection Type and Visual Clutter for Spatial Reference. In C. Freksa, B. Nebel, M. Hegarty, T. Barkowsky (Eds.), *Report Series of the Transregional Collaborative Research Center SFB/TR 8 Spatial Cognition.* Bremen: University of Bremen.

Baltaretu, A., Schilperoord, J., & Salami, G. (2014). Power metaphor as size difference. In W. Hollmann, D. Van Olmen, S. B. C. Hart (Eds.), *roceedings of the 5th UK Cognitive Linguistics Conference.* Lancaster: Lancaster University.

Baltaretu, A., Krahmer, E., & Maes, A. (2013). Factors influencing the choice of relatum in referring expressions generation: animacy vs. position. In A. Gatt, R. van Gompel, E. G. Bard, E. Krahmer, K. van Deemter (Eds.), *Proceedings of the CogSci workshop on the production of referring expressions: bridging the gap between cognitive and computational approaches to reference.* pp. 1 - 6. Berlin: PRE-CogSci.

Baltaretu, A., Maes, A. & van Wijk, C. (2012). Mere presence, object orientation and perspective taking. In E. de Vries, & K. Scheiter (Eds.), *Proceedings of the Bi-Annual Conference of EARLI Special Interest Group Comprehension of Text and Graphics, Staging knowledge and experience.* pp. 28 - 30. Grenoble: University Pierre-Mendes.

TiCC Ph.D. Series

1. Pashiera Barkhuysen. *Audiovisual Prosody in Interaction.* Promotores: M.G.J. Swerts, E.J. Krahmer. Tilburg, 3 October 2008.

2. Ben Torben-Nielsen. *Dendritic Morphology: Function Shapes Structure.* Promotores: H.J. van den Herik, E.O. Postma. Copromotor: K.P. Tuyls. Tilburg, 3 December 2008.

3. Hans Stol. *A Framework for Evidence-based Policy Making Using IT.* Promotor: H.J. van den Herik. Tilburg, 21 January 2009.

4. Jeroen Geertzen. *Dialogue Act Recognition and Prediction.* Promotor: H. Bunt. Copromotor: J.M.B. Terken. Tilburg, 11 February 2009.

5. Sander Canisius. *Structured Prediction for Natural Language Processing.* Promotores: A.P.J. van den Bosch, W. Daelemans. Tilburg, 13 February 2009.

6. Fritz Reul. *New Architectures in Computer Chess.* Promotor: H.J. van den Herik. Copromotor: J.W.H.M. Uiterwijk. Tilburg, 17 June 2009.

7. Laurens van der Maaten. *Feature Extraction from Visual Data.* Promotores: E.O. Postma, H.J. van den Herik. Copromotor: A.G. Lange. Tilburg, 23 June 2009 (cum laude).

8. Stephan Raaijmakers. *Multinomial Language Learning.* Promotores: W. Daelemans, A.P.J. van den Bosch. Tilburg, 1 December 2009.

9. Igor Berezhnoy. *Digital Analysis of Paintings.* Promotores: E.O. Postma, H.J. van den Herik. Tilburg, 7 December 2009.

10. Toine Bogers. *Recommender Systems for Social Bookmarking.* Promotor: A.P.J. van den Bosch. Tilburg, 8 December 2009.

11. Sander Bakkes. *Rapid Adaptation of Video Game AI.* Promotor: H.J. van den Herik. Copromotor: P. Spronck. Tilburg, 3 March 2010.

12. Maria Mos. *Complex Lexical Items.* Promotor: A.P.J. van den Bosch. Copromotores: A. Vermeer, A. Backus. Tilburg, 12 May 2010 (in collaboration with the Department of Language and Culture Studies).

13. Marieke van Erp. *Accessing Natural History. Discoveries in data cleaning, structuring, and retrieval.* Promotor: A.P.J. van den Bosch. Copromotor: P.K. Lendvai. Tilburg, 30 June 2010.

14. Edwin Commandeur. *Implicit Causality and Implicit Consequentiality in Language Comprehension.* Promotores: L.G.M. Noordman, W. Vonk. Copromotor: R. Cozijn. Tilburg, 30 June 2010.

15. Bart Bogaert. *Cloud Content Contention.* Promotores: H.J. van den Herik, E.O. Postma. Tilburg, 30 March 2011.

16. Xiaoyu Mao. *Airport under Control.* Promotores: H.J. van den Herik, E.O. Postma. Copromotores: N. Roos, A. Salden. Tilburg, 25 May 2011.

17. Olga Petukhova. *Multidimensional Dialogue Modelling.* Promotor: H. Bunt. Tilburg, 1 September 2011.

18. Lisette Mol. *Language in the Hands.* Promotores: E.J. Krahmer, A.A. Maes, M.G.J. Swerts. Tilburg, 7 November 2011 (cum laude).

19. Herman Stehouwer. *Statistical Language Models for Alternative Sequence Selection.* Promotores: A.P.J. van den Bosch, H.J. van den Herik. Copromotor: M.M. van Zaanen. Tilburg, 7 December 2011.

20. Terry Kakeeto-Aelen. *Relationship Marketing for SMEs in Uganda.* Promotores: J. Chr. van Dalen, H.J. van den Herik. Copromotor: B.A. Van de Walle. Tilburg, 1 February 2012.

21. Suleman Shahid. *Fun & Face: Exploring non-verbal expressions of emotion during playful interactions.* Promotores: E.J. Krahmer, M.G.J. Swerts. Tilburg, 25 May 2012.

22. Thijs Vis. *Intelligence, Politie en Veiligheidsdienst: Verenigbare Grootheden?* Promotores: T.A. de Roos, H.J. van den Herik, A.C.M. Spapens. Tilburg, 6 June 2012 (in collaboration with the Tilburg School of Law).

23. Nancy Pascall. *Engendering Technology Empowering Women.* Promotores: H.J. van den Herik, M. Diocaretz. Tilburg, 19 November 2012.

24. Agus Gunawan. *Information Access for SMEs in Indonesia.* Promotor: H.J. van den Herik. Copromotores: M. Wahdan, B.A. Van de Walle. Tilburg, 19 December 2012.

25. Giel van Lankveld. *Quantifying Individual Player Differences.* Promotores: H.J. van den Herik, A.R. Arntz. Copromotor: P. Spronck. Tilburg, 27 February 2013.

26. Sander Wubben. *Text-to-text Generation Using Monolingual Machine Translation.* Promotores: E.J. Krahmer, A.P.J. van den Bosch, H. Bunt. Tilburg, 5 June 2013.

27. Jeroen Janssens. *Outlier Selection and One-Class Classification.* Promotores: E.O. Postma, H.J. van den Herik. Tilburg, 11 June 2013.

28. Martijn Balsters. *Expression and Perception of Emotions: The Case of Depression, Sadness and Fear.* Promotores: E.J. Krahmer, M.G.J. Swerts, A.J.J.M. Vingerhoets. Tilburg, 25 June 2013.

29. Lisanne van Weelden. *Metaphor in Good Shape.* Promotor: A.A. Maes. Copromotor: J. Schilperoord. Tilburg, 28 June 2013.

30. Ruud Koolen. *Need I say More? On Overspecification in Definite Reference.* Promotores: E.J. Krahmer, M.G.J. Swerts. Tilburg, 20 September 2013.

31. J. Douglas Mastin. *Exploring Infant Engagement. Language Socialization and Vocabulary Development: A Study of Rural and Urban Communities in Mozambique.* Promotor: A.A. Maes. Copromotor: P.A. Vogt. Tilburg, 11 October 2013.

32. Philip C. Jackson. Jr. *Toward Human-Level Artificial Intelligence – Representation and Computation of Meaning in Natural Language.* Promotores: H.C. Bunt, W.P.M. Daelemans. Tilburg, 22 April 2014.

33. Jorrig Vogels. *Referential choices in language production: The Role of Accessibility.* Promotores: A.A. Maes, E.J. Krahmer. Tilburg, 23 April 2014.

34. Peter de Kock. *Anticipating Criminal Behaviour.* Promotores: H.J. van den Herik, J.C. Scholtes. Copromotor: P. Spronck. Tilburg, 10 September 2014.

35. Constantijn Kaland. *Prosodic marking of semantic contrasts: do speakers adapt to addressees?* Promotores: M.G.J. Swerts, E.J. Krahmer. Tilburg, 1 October 2014.

36. Jasmina Marič. *Web Communities, Immigration and Social Capital.* Promotor: H.J. van den Herik. Copromotores: R. Cozijn, M. Spotti. Tilburg, 18 November 2014.

37. Pauline Meesters. *Intelligent Blauw.* Promotores: H.J. van den Herik, T.A. de Roos. Tilburg, 1 December 2014.

38. Mandy Visser. *Better use your head.* How people learn to signal emotions in social contexts. Promotores: M.G.J. Swerts, E.J. Krahmer. Tilburg, 10 June 2015.

39. Sterling Hutchinson. *How symbolic and embodied representations work in concert.* Promotores: M.M. Louwerse, E.O. Postma. Tilburg, 30 June 2015.

40. Marieke Hoetjes. *Talking hands. Reference in speech, gesture and sign.* Promotores: E.J. Krahmer, M.G.J. Swerts. Tilburg, 7 October 2015

41. Elisabeth Lubinga. *Stop HIV. Start talking? The effects of rhetorical figures in health messages on conversations among South African adolescents.* Promotores: A.A. Maes, C.J.M. Jansen. Tilburg, 16 October 2015.

42. Janet Bagorogoza. *Knowledge Management and High Performance. The Uganda Financial Institutions Models for HPO.* Promotores: H.J. van den Herik, B. van der Walle, Tilburg, 24 November 2015.

43. Hans Westerbeek. *Visual realism: Exploring effects on memory, language production, comprehension, and preference.* Promotores: A.A. Maes, M.G.J. Swerts. Copromotor: M.A.A. van Amelsvoort. Tilburg, 10 Februari 2016.

44. Matje van de Camp. *A link to the Past: Constructing Historical Social Networks from Unstructured Data.* Promotores: A.P.J. van den Bosch, E.O. Postma. Tilburg, 2 Maart 2016.

45. Annemarie Quispel. *Data for all: Data for all: How professionals and non-professionals in design use and evaluate information visualizations.* Promotor: A.A. Maes. Copromotor: J. Schilperoord. Tilburg, 15 Juni 2016.

46. Rick Tillman. *Language Matters: The Influence of Language and Language Use on Cognition.* Promotores: M.M. Louwerse, E.O. Postma. Tilburg, 30 Juni 2016.

47. Ruud Mattheij. *The Eyes Have It.* Promoteres: E.O. Postma, H. J. Van den Herik, and P.H.M. Spronck. Tilburg, 5 October 2016.

48. Marten Pijl. *Tracking of human motion over time.* Promotores: E. H. L. Aarts, M. M. Louwerse. Copromotor: J. H. M. Korst. Tilburg, 14 December 2016.

49. Yevgen Matusevych. *Learning constructions from bilingual exposure: Computational studies of argument structure acquisition.* Promotor: A.M. Backus. Copromotor: A. Alishahi. Tilburg 19 December 2016.

50. Karin van Nispen. *What can people with aphasia communicate with their hands? A study of representation techniques in pantomime and co-speech gesture.* Promotor: E.J. Krahmer. Tilburg, 19 December 2016.

51. Adriana Baltaretu. *Speaking of landmarks. How visual information influences reference in spatial domains.* Promotores: A.A. Maes and E.J. Krahmer. Tilburg, 22 december 2016.

Bibliography

Abrams, R. A., & Christ, S. E. (2003). Motion onset captures attention. *Psychological Science*, *14*(5), 427–432. doi: 10.1111/1467-9280.01458

Agrawala, M., & Stolte, C. (2001). Rendering effective route maps: improving usability through generalization. In L. Pocock (Ed.), *Proceedings of the 28th annual conference on Computer Graphics and Interactive Techniques* (pp. 241–249). Los Angeles, CA: ACM Special Interest Group on Graphics and Interactive Techniques.

Allen, G. L. (2000). Principles and practices for communicating route knowledge. *Applied Cognitive Psychology*, *14*(4), 333–359. doi: 10.1002/1099-0720

Allen, G. L., Siegel, A. W., & Rosinski, R. R. (1978). The role of perceptual context in structuring spatial knowledge. *Journal of Experimental Psychology: Human Learning and Memory*, *4*(6), 617–630. doi: 10.1037/0278-7393.4.6.617

Ariel, M. (1990). *Accessing noun-phrase antecedents*. London: Routlege.

Arnold, J. E., & Lao, S.-Y. C. (2015). Effects of psychological attention on pronoun comprehension. *Language, Cognition and Neuroscience*, *30*(7), 832–852.

Arts, A., Maes, A., Noordman, L., & Jansen, C. (2011). Overspecification facilitates object identification. *Journal of Pragmatics*, *43*(1), 361–374.

Asher, M., Tolhurst, D., Troscianko, T., & Gilchrist, I. (2013). Regional effects of clutter on human target detection performance. *Journal of Vision*, *13*(5), 25–25.

Bangerter, A. (2004). Using pointing and describing to achieve joint focus of attention in dialogue. *Psychological Science*, *15*(6), 415–419.

Barclay, M., & Galton, A. (2008). An influence model for reference object selection in spatially locative phrases. In C. Freksa, N. Newcombe, P. Gärdenfors, &

S. Wölfl (Eds.), *Spatial Cognition VI. Learning, Reasoning, and Talking about Space* (pp. 216–232). Berlin: Springer.

Barclay, M., & Galton, A. (2013). Selection of reference objects for locative expressions: the importance of knowledge and perception. In T. Tenbrink, J. Wiener, & C. Claramunt (Eds.), *Representing space in cognition: Interrelations o behavior, langauge, and formal models* (pp. 57–169). Oxford: Oxford University Press.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Belke, E., & Meyer, A. S. (2002). Tracking the time course of multidimensional stimulus discrimination: Analyses of viewing patterns and processing times during "same"-"different" decisions. *European Journal of Cognitive Psychology*, *14*(2), 237–266.

Beun, R.-J., & Cremers, A. (1998). Object reference in a shared domain of conversation. *Pragmatics & Cognition*, *6*(1-2), 121–152.

Bock, K., Irwin, D., Davidson, D., & Levelt, W. (2003). Minding the clock. *Journal of Memory and Language*, *48*(4), 653–685.

Bock, K., Loebell, H., & Morey, R. (1992). From conceptual roles to structural relations: bridging the syntactic cleft. *Psychological Review*, *99*(1), 150.

Bock, K., & Warren, R. (1985). Conceptual accessibility and syntactic structure in sentence formulation. *Cognition*, *21*(1), 47–67.

Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *35*(1), 185–207. doi: 10.1109/TPAMI.2012.89

Boroditsky, L. (2000). Metaphoric structuring: Understanding time through spatial metaphors. *Cognition*, *75*(1), 1–28.

Branigan, H. P., Pickering, M. J., & Tanaka, M. (2008). Contributions of animacy to grammatical function assignment and word order during production. *Lingua*, *118*(2), 172–189.

Bravo, M., & Farid, H. (2004). Recognizing and segmenting objects in clutter. *Vision Research*, *44*(4), 385–396.

Bravo, M., & Farid, H. (2006). Object recognition in dense clutter. *Perception & Psychophysics*, *68*(6), 911–918.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1482-1493.

Brennan, S. E., Schuhmann, K. S., & Batres, K. M. (2013). Entrainment on the

move and in the lab: The Walking Around Corpus. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 1934–1939). Austin, TX: Cognitive Science Society.

Bresnan, J., Cueni, A., Nikitina, T., & Baayen, H. (2007). Predicting the dative alternation. *Cognitive Foundations of Interpretation*, 69–94.

Brown-Schmidt, S., & Tanenhaus, M. (2006). Watching the eyes when talking about size: An investigation of message formulation and utterance planning. *Journal of Memory and Language*, *54*(4), 592–609.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data? *Perspectives on psychological science*, *6*(1), 3–5. doi: 10.1177/1745691610393980

Burgess, N., Spiers, H. J., & Paleologou, E. (2004). Orientational manoeuvres in the dark: dissociating allocentric and egocentric influences on spatial memory. *Cognition*, *94*(2), 149–166. doi: 10.1016/j.cognition.2004.01.001

Butler, L. K., Tilbe, T. J., Jaeger, T. F., & Bohnemeyer, J. (2014). Order of nominal conjuncts in visual scene description depends on language. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th annual conference of the Cognitive Science Society.* Cognitive Science Society: Austin, TX.

Byron, D. K., & Fosler-Lussier, E. (2006). The OSU Quake 2004 corpus of two-party situated problem-solving dialogs. In N. Calzolari (Ed.), *Proceedings of the 15th Language Resources and Evaluation Conference.* European Language Resources Association.

Campbell, J. (1993). The role of physical objects in spatial thinking. In N. Eilan, R. McCarthy, & B. Brewer (Eds.), *Spatial Representation* (pp. 65–95). Malden: Blackwell Publishing.

Carlson-Radvansky, L., Covey, E., & Lattanzi, K. (1999). "What" effects on "where": Functional influences on spatial relations. *Psychological Science*, *10*(6), 516–521.

Carlson-Radvansky, L., & Radvansky, G. (1996). The influence of functional relations on spatial term selection. *Psychological Science*, *7*(1), 56–60.

Carmi, R., & Itti, L. (2006). Visual causes versus correlates of attentional selection in dynamic scenes. *Vision Research*, *46*(26), 4333–4345. doi: 10.1016/j.visres.2006.08.019

Casasanto, D., & Boroditsky, L. (2008). Time in the mind: Using space to think about time. *Cognition*, *106*(2), 579–593.

Chan, E., Baumann, O., Bellgrove, M. A., & Mattingley, J. B. (2012). From objects to landmarks: the function of visual location information in spatial navigation. *Frontiers in Psychology*, *3*, 304. doi: 10.3389/fpsyg.2012.00304

Chan, T. T., & Bergen, B. (2005). Writing direction influences spatial cognition.

In B. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th annual conference of the Cognitive Science Society* (pp. 412–417). Austin, TX: Cognitive Science Society.

Chatterjee, A. (2001). Language and space: Some interactions. *Trends in Cognitive Sciences*, *5*(2), 55–61.

Chokron, S., & De Agostini, M. (2000). Reading habits influence aesthetic preference. *Cognitive Brain Research*, *10*(1), 45–49.

Chokron, S., & Imbert, M. (1993). Influence of reading habits on line bisection. *Cognitive Brain Research*, *1*(4), 219–222.

Clark, H. H. (1996). *Using language*. Cambridge, England: Cambridge University Press.

Clark, H. H., & Bangerter, A. (2004). Changing ideas about reference. In *Experimental pragmatics* (pp. 25–49). Springer.

Clark, H. H., & Begun, J. S. (1971). The semantics of sentence subjects. *Language and Speech*, *14*(1), 34–46.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39.

Clarke, A., Coco, M., & Keller, F. (2013). The impact of attentional, linguistic and visual features during object naming. *Frontiers in Psychology*, *4*(927). doi: 10.3389/fpsyg.2013.00927

Clarke, A., Elsner, M., & Rohde, H. (2013). Where's Wally: the influence of visual salience on referring expression generation. *Frontiers in Psychology*, *4*, 329. doi: 10.3389/fpsyg.2013.00329

Clarke, A., Elsner, M., & Rohde, H. (2015). Giving good directions: order of mention reflects visual salience. *Frontiers in Psychology*, *6*(1793). doi: 10.3389/fpsyg.2015.01793

Clarke, A., & Tatler, B. (2014). Deriving an appropriate baseline for describing fixation behaviour. *Vision Research*, *102*, 41–51. doi: 10.1016/j.visres.2014.06.016

Coco, M. I., & Keller, F. (2009). The impact of visual information on reference assignment in sentence production. In N. Taatgen, H. van Rijn, L. Schomaker, & J. Nerbonne (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 274–279). Austin, TX: Cognitive Science Society.

Coco, M. I., & Keller, F. (2015). Integrating mechanisms of visual guidance in naturalistic language production. *Cognitive Processing*, *16*(2), 131–150.

Couclelis, H. (1996). Verbal directions for way-finding: Space, cognition, and language. In J. Portugali (Ed.), *The construction of cognitive maps* (pp. 133–153). Berlin: Springer.

Couclelis, H., Golledge, R. G., Gale, N., & Tobler, W. (1987). Exploring the anchor-point hypothesis of spatial cognition. *Journal of Environmental Psychology*, *7*(2), 99–122.

Coventry, K. R., & Garrod, S. C. (2004). *Saying, seeing and acting: The psychological semantics of spatial prepositions.* NY: Psychology Press.

Craton, L. G., Elicker, J., Plumert, J. M., & Pick, H. L. (1990). Children's use of frames of reference in communication of spatial location. *Child Development*, *61*(5), 1528–1543.

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PLoS ONE*, *8*(3), e57410. doi: 10.1371/journal.pone.0057410

Dale, R., Geldof, S., & Prost, J.-P. (2005). Using natural language generation in automatic route. *Journal of Research and practice in Information Technology*, *37*(1), 89.

Dale, R., & Haddock, N. (1991). Generating referring expressions involving relations. In J. Kunze & D. Reimann (Eds.), *Proceedings of the fifth conference of the European Association for Computational Linguistics* (pp. 161–166). Stroudsburg, PA: Association for Computational Linguistics.

Dale, R., & Reiter, E. (1995). Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, *19*(2), 233–263.

Dale, R., & Viethen, J. (2009). Referring expression generation through attribute-based heuristics. In E. Krahmer & M. Theune (Eds.), *Proceedings of the 12th European Workshop on Natural Language Generation* (pp. 58–65). Stroudsburg, PA: Association for Computational Linguistics.

Daniel, M.-P., & Denis, M. (2004). The production of route directions: Investigating conditions that favour conciseness in spatial discourse. *Applied Cognitive Psychology*, *18*(1), 57–75.

Davies, C., & Katsos, N. (2009). Are interlocutors as sensitive to over-informativeness as they are to under-informativeness. In K. van Deemter, A. Gatt, R. van Gompel, & E. Krahmer (Eds.), *Proceedings of the Pre-CogSci workshop on the Production of Referring Expressions: Bridging the gap between computational and empirical approaches to reference.* Amsterdam: Pre-CogSci.

Davies, C., & Katsos, N. (2013). Are speakers and listeners 'only moderately Gricean'? An empirical response to Engelhardt et al.(2006). *Journal of Pragmatics*, *49*(1), 78–106.

Denis, M., Mores, C., Gras, D., Gyselinck, V., & Daniel, M.-P. (2014). Is memory for routes enhanced by an environment's richness in visual landmarks? *Spatial Cognition & Computation*, *14*(4), 284–305.

Denis, M., Pazzaglia, F., Cornoldi, C., & Bertolo, L. (1999). Spatial discourse and navigation: An analysis of route directions in the city of Venice. *Applied Cognitive Psychology*, *13*(2), 145–174. doi: 10.1002/(SICI)1099-0720(199904)13:2⟨145::AID-ACP550⟩3.0.CO;2-4

de Vega, M., Rodrigo, M. J., Ato, M., Dehn, D. M., & Barquero, B. (2002). How nouns and prepositions fit together: An exploration of the semantics of locative

sentences. *Discourse Processes*, *34*(2), 117–143.

Dickinson, C. A., & Intraub, H. (2009). Spatial asymmetries in viewing and remembering scenes: Consequences of an attentional bias? *Attention, Perception, & Psychophysics*, *71*(6), 1251–1262.

Di Eugenio, B., Jordan, P. W., Thomason, R. H., & Moore, J. (2000). The agreement process: An empirical investigation of human–human computer-mediated collaborative dialogs. *International Journal of Human-Computer Studies*, *53*(6), 1017–1076.

Dollar, P., Wojek, C., Schiele, B., & Perona, P. (2012). Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(4), 743–761. doi: 10.1109/TPAMI.2011.155

Donderi, D., & McFadden, S. (2005). Compressed file length predicts search time and errors on visual displays. *Displays*, *26*(2), 71–78.

Dos Santos Silva, D., & Paraboni, I. (2015). Generating Spatial Referring Expressions in Interactive 3D Worlds. *Spatial Cognition & Computation*, *15*(3), 186-225. doi: 10.1080/13875868.2015.1039166

Downing, P. E., Bray, D., Rogers, J., & Childs, C. (2004). Bodies capture attention when nothing is expected. *Cognition*, *93*(1), 27–38.

Duckham, M., Winter, S., & Robinson, M. (2010). Including landmarks in routing instructions. *Journal of Location Based Services*, *4*(1), 28–52. doi: 10.1080/17489721003785602

D'Zmura, M. (1991). Color in visual search. *Vision Research*, *31*(6), 951–966.

Einhäuser, W., Rutishauser, U., & Koch, C. (2008). Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of Vision*, *8*(2), 1–19. doi: 10.1167/8.2.2

Elias, B., Paelke, V., & Kuhnt, S. (2005). Concepts for the cartographic visualization of landmarks. In G. Gartner, W. Cartwright, & M. P. Peterson (Eds.), *Location based services & telecartography-proceedings of the symposium* (pp. 1149–1155). Berlin Heidelberg: Springer-Verlag.

Elsner, M., Rohde, H., & Clarke, A. (2014). Information Structure Prediction for Visual-world Referring Expressions. In S. Wintner, S. Goldwater, & S. Riezler (Eds.), *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 520–530). Stroudsburg, PA: Association for Computational Linguistics.

Engelhardt, P. E., Bailey, K. G., & Ferreira, F. (2006). Do speakers and listeners observe the gricean maxim of quantity? *Journal of Memory and Language*, *54*(4), 554–573.

Feng, Y., & Lapata, M. (2010). How many words is a picture worth? automatic caption generation for news images. In J. Hajič (Ed.), *Proceedings of the 48th annual meeting of the Association for Computational Linguistics* (pp. 1239–1249). Stroudsburg, PA: Association for Computational Linguistics.

Ferguson, E. L., & Hegarty, M. (1994). Properties of cognitive maps constructed from texts. *Memory & Cognition*, *22*(4), 455–473. doi: 10.3758/BF03200870

Fletcher-Watson, S., Findlay, J., Leekam, S., & Benson, V. (2008). Rapid detection of person information in a naturalistic scene. *Perception*, *37*(4), 571–583.

Folk, C. L., Remington, R. W., & Wright, J. H. (1994). The structure of attentional control: contingent attentional capture by apparent motion, abrupt onset, and color. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(2), 317-329.

Forrest, L. B. (1996). Discourse goals and attentional processes in sentence production: The dynamic construal of events. In A. Goldberg (Ed.), *Conceptual structure, discourse and language* (pp. 149–161). Stanford, CA: Center for the Study of Language and Information.

Foulsham, T., Gray, A., Nasiopoulos, E., & Kingstone, A. (2013). Leftward biases in picture scanning and line bisection: A gaze-contingent window study. *Vision Research*, *78*, 14–25. doi: 10.1016/j.visres.2012.12.001

Frank, M., & Goodman, N. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Fraurud, K. (1996). Cognitive ontology and NP form. *Pragmatics and Beyond New Series*, 65–88.

Fukumura, K., & van Gompel, R. (2011). The effect of animacy on the choice of referring expression. *Language and Cognitive Processes*, *26*(10), 1472–1504.

Gardent, C. (2002). Generating minimal definite descriptions. In J. Stephen (Ed.), *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (pp. 96–103). Stroudsburg, PA: Association for Computational Linguistics.

Gargett, A., Garoufi, K., Koller, A., & Striegnitz, K. (2010). The GIVE-2 corpus of generating instructions in virtual environments. In N. Calzolari, K. Choukri, & B. Maegaard (Eds.), *Proceedings of the 7th International Conference on Language Resources and Evaluation.* Malta: European Language Resources Association.

Garoufi, K. (2013). *Interactive generation of effective discourse in situated context: A planning-based approach.* Unpublished doctoral dissertation, University of Postdam.

Garoufi, K., & Koller, A. (2010). Automated planning for situated natural language generation. In J. Hajič (Ed.), *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 1573–1582). Stroudsburg, PA: Association for Computational Linguistics.

Garoufi, K., Staudte, M., Koller, A., & Crocker, M. W. (2015). Exploiting listener gaze to improve situated communication in dynamic virtual environments. *Cognitive Science*, *early view*.

Gatt, A., & Belz, A. (2010). Introducing shared tasks to NLG: The TUNA shared task

evaluation challenges. In E. Krahmer & M. Theune (Eds.), *Empirical methods in natural language generation* (pp. 264–293). Berlin: Springer.

Gatt, A., Krahmer, E., van Deemter, K., & van Gompel, R. (2014). Models and empirical data for the production of referring expressions. *Language, Cognition and Neuroscience*, *29*(8), 899–911.

Gentner, D., Özyürek, A., Gürcanli, Ö., & Goldin-Meadow, S. (2013). Spatial language facilitates spatial cognition: Evidence from children who lack language input. *Cognition*, *127*(3), 318–330.

Gkatzia, D., Rieser, V., Bartie, P., & Mackaness, W. (2015). From the Virtual to the Real World: Referring to Objects in Real-World Spatial Scenes. In L. Marquez, C. Callison-Burch, & J. Su (Eds.), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 1936–1942). Stroudsburg, PA: Association for Computational Linguistics.

Gleitman, L. R., January, D., Nappa, R., & Trueswell, J. C. (2007). On the give and take between event apprehension and utterance formulation. *Journal of Memory and Language*, *57*(4), 544–569.

Golding, J. M., Graesser, A. C., & Hauselt, J. (1996). The processing of answering direction-giving questions when someone is lost on a university campus: The role of pragmatics. *Applied Cognitive Psychology*, *10*, 23–39.

Goudbeek, M., & Krahmer, E. (2012). Alignment in interactive reference production: Content planning, modifier ordering, and referential overspecification. *Topics in Cognitive Science*, *4*(2), 269–289. doi: 10.1111/j.1756-8765.2012.01186.x

Grice, H. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Speech acts* (pp. 41–58). New York: Academic Press.

Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science*, *11*(4), 274–279.

Hanna, J., & Brennan, S. E. (2007). Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language*, *57*(4), 596–615.

Henderson, J., Chanceaux, M., & Smith, T. (2009). The influence of clutter on real-world scene search: Evidence from search efficiency and eye movements. *Journal of Vision*, *9*(1), 32–32.

Henderson, J., & Ferreira, F. (2013). *The interface of language, vision, and action: Eye movements and the visual world*. New York: Psychology Press.

Hillstrom, A. P., & Yantis, S. (1994). Visual motion and attentional capture. *Perception & Psychophysics*, *55*(4), 399–411.

Hirtle, S., Richter, K.-F., Srinivas, S., & Firth, R. (2010). This is the tricky part: When directions become difficult. *Journal of Spatial Information Science*, *2010*(1), 53–73.

Hölscher, C., Tenbrink, T., & Wiener, J. M. (2011). Would you follow your own route description? cognitive strategies in urban route planning. *Cognition*, *121*(2),

228–247.

Horacek, H. (1997). An algorithm for generating referential descriptions with flexible interfaces. In P. R. Cohen & W. Wahlster (Eds.), *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics* (pp. 206–213). Stroudsburg, PA: Association for Computational Linguistics.

Hund, A. M., & Plumert, J. M. (2007). What counts as by? Young children's use of relative distance to judge nearbyness. *Developmental Psychology*, *43*(1), 121.

Ishikawa, T., & Montello, D. R. (2006). Spatial knowledge acquisition from direct experience in the environment: Individual differences in the development of metric knowledge and the integration of separately learned places. *Cognitive Psychology*, *52*(2), 93–129.

Ishikawa, T., & Nakamura, U. (2012). Landmark selection in the environment: Relationships with object characteristics and sense of direction. *Spatial Cognition & Computation*, *12*(1), 1–22. doi: 10.1080/13875868.2011.581773

Itti, L. (2005). Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Visual Cognition*, *12*(6), 1093–1123. doi: 10.1080/13506280444000661

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*(3), 194–203.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446.

Jaeger, T. F., & Tily, H. (2011). On language 'utility': Processing complexity and communicative efficiency. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(3), 323–335.

Janarthanam, S., Lemon, O., Bartie, P., Dalmas, T., Dickinson, A., Liu, X., ... Webber, B. (2013). Evaluating a city exploration dialogue system combining question-answering and pedestrian navigation. In H. Schuetze (Ed.), *The 51st Annual Meeting of the Association for Computational Linguistics* (pp. 1660–1668). Stroudsburg, PA: The Association for Computational Linguistics.

Jordan, P. W., & Walker, M. A. (2005). Learning content selection rules for generating object descriptions in dialogue. *Journal of Artificial Intelligence Research*, *24*, 157–194.

Kazemzadeh, S., Ordonez, V., Matten, M., & Berg, T. (2014). ReferItGame: Referring to objects in photographs of natural scenes. In A. Moschitti (Ed.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 787–798). Stroudsburg, PA: Association for Computational Linguistics.

Kelleher, J., Costello, F., & van Genabith, J. (2005). Dynamically structuring, updating and interrelating representations of visual and linguistic discourse context.

Artificial Intelligence, *167*(1), 62–102.

Kelleher, J., & Kruijff, G.-J. (2006). Incremental generation of spatial referring expressions in situated dialog. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics* (pp. 1041–1048). Stroudsburg, PA: Association for Computational Linguistics.

Kelleher, J., Ross, R., Mac Namee, B., & Sloan, C. (2010). Situating spatial templates for human-robot interaction. In R. Pirrone, R. Azevedo, & G. Biswas (Eds.), *Aaai fall symposium series.* Arlington, Virginia: American Association for Artificial Intelligence Fall Symposium Series.

Kirchner, H., & Thorpe, S. J. (2006). Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. *Vision research*, *46*(11), 1762–1776.

Kirsh, D. (1995). The intelligent use of space. *Artificial Intelligence*, *73*(1), 31–68.

Klippel, A. (2003). Wayfinding choremes. In W. Kuhn, M. Worboys, & S. Timpf (Eds.), *International Conference on Spatial Information Theory* (pp. 301–315). Berlin: Springer.

Klippel, A., Tenbrink, T., & Montello, D. R. (2013). The role of structure and function in the conceptualization of directions. In M. Vulchanova & E. van der Zee (Eds.), *Motion encoding in language and space* (pp. 102–199). Oxford: Oxford University Press.

Klippel, A., & Winter, S. (2005). Structural salience of landmarks for route directions. In A. Cohn & D. Mark (Eds.), *International conference on spatial information theory* (pp. 347–362). Berlin Heidelberg: Springer.

Kollmorgen, S., Nortmann, N., Schröder, S., & König, P. (2010). Influence of low-level stimulus features, task dependent factors, and spatial biases on overt visual attention. *PLoS Computational Biology*, *6*(5), e1000791. doi: 10.1371/journal .pcbi.1000791

Koolen, R., Gatt, A., Goudbeek, M., & Krahmer, E. (2011). Factors causing overspecification in definite descriptions. *Journal of Pragmatics*, *43*(13), 3231–3250.

Koolen, R., Krahmer, E., & Swerts, M. (2013). The impact of bottom-up and top-down saliency cues on reference production. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th annual meeting of the Cognitive Science Society* (pp. 817–822). Austin, TX: Cognitive Science Society.

Koolen, R., Krahmer, E., & Theune, M. (2012). Learning preferences for referring expression generation: Effects of domain, language and algorithm. In B. Di Eugenio & S. McRoy (Eds.), *Proceedings of the Seventh International Natural Language Generation Conference* (pp. 3–11). Stroudsburg, PA: Association for Computational Linguistics.

Krahmer, E., & Theune, M. (2002). Efficient context-sensitive generation of referring expressions. In K. van Deemter & R. Kibble (Eds.), *Information sharing:*

Reference and presupposition in language generation and interpretation (pp. 223–263). Stanford: CSLI Publications.

Krahmer, E., & van Deemter, K. (2012). Computational generation of referring expressions: A survey. *Computational Linguistics*, *38*(1), 173–218.

Krahmer, E., Van Erk, S., & Verleg, A. (2003). Graph-based generation of referring expressions. *Computational Linguistics*, *29*(1), 53–72.

Krieg-Brückner, B. (1998). A taxonomy of spatial knowledge for navigation. In U. Schmid & F. Wysotzki (Eds.), *Qualitative and quantitative approaches to spatial inference and the analysis of movements* (p. 98-102). Berlin: Technische Universität Berlin.

Lakoff, G. (1975). *Hedges: a study in meaning criteria and the logic of fuzzy concepts.* Berlin: Springer.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by.* Chicago: University of Chicago Press.

Learmonth, A. E., Newcombe, N. S., & Huttenlocher, J. (2001). Toddlers' use of metric information and landmarks to reorient. *Journal of Experimental Child Psychology*, *80*(3), 225–244.

Lee, P., Tappe, H., & Klippel, A. (2002). Acquisition of landmark knowledge from static and dynamic presentation of route maps. In W. D. Gray & C. Schunn (Eds.), *Proceedings of the 24th Annual Conference of the Cognitive Science Society* (pp. 1017 – 1018). Austin, TX: Cognitive Science Society.

Lee, P., & Tversky, B. (2005). Interplay between visual and spatial: The effect of landmark descriptions on comprehension of route/survey spatial descriptions. *Spatial Cognition & Computation*, *5*(2-3), 163–185. doi: 10.1080/13875868.2005.9683802

Levelt, W. J. (1993). *Speaking: From intention to articulation.* Cambridge: MIT press.

Levinson, S. C. (1996). Frames of reference and Molyneux's question: Crosslinguistic evidence. , 109–169.

Levinson, S. C. (2003). *Space in language and cognition: Explorations in cognitive diversity.* Cambridge University Press.

Louwerse, M., Benesh, N., Hoque, M., Jeuniaux, P., Lewis, G., Wu, J., & Zirnstein, M. (2007). Multimodal communication in face-to-face conversations. In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Cognitive Science Society* (pp. 1235–1240). Austin, TX: Cognitive Science Society.

Lovelace, K. L., Hegarty, M., & Montello, D. R. (1999). Elements of good route directions in familiar and unfamiliar environments. In C. Freksa & D. M. Mark (Eds.), *Spatial information theory. Cognitive and computational foundations of geographic information science* (pp. 65–82). Berlin: Springer.

Maass, A., & Russo, A. (2003). Directional bias in the mental representation of spatial events nature or culture? *Psychological Science*, *14*(4), 296–301.

Maes, A., Arts, A., & Noordman, L. (2004). Reference management in instructive discourse. *Discourse Processes*, *37*(2), 117–144.

Mast, V., Couto Vale, D., & Falomir, Z. (2014). Enabling grounding dialogues through probabilistic reference handling. In A. Eshghi, K. Fukumura, & S. Janarthanam (Eds.), *Proceedings of RefNet Workshop on Psychological and Computational Models of Reference Comprehension and Production.* Edinburgh: RefNet.

Mast, V., Smeddinck, J., Strotseva, A., & Tenbrink, T. (2010). The impact of dimensionality on natural language route directions in unconstrained dialogue. In R. Fern'/andez & O. Lemon (Eds.), *Proceedings of the 11th annual meeting of the special interest group on discourse and dialogue* (pp. 99–102). Stroudsburg, PA: Association for Computational Linguistics.

May, A. J., Ross, T., Bayer, S. H., & Tarkiainen, M. J. (2003). Pedestrian navigation aids: information requirements and design implications. *Personal and Ubiquitous Computing*, *7*(6), 331–338. doi: 10.1007/s00779-003-0248-5

McDonald, J., Bock, K., & Kelly, M. (1993). Word and world order: Semantic, phonological, and metrical determinants of serial position. *Cognitive Psychology*, *25*(2), 188–230.

Meyer, A. S., Sleiderink, A. M., & Levelt, W. J. (1998). Viewing and naming objects: Eye movements during noun phrase production. *Cognition*, *66*(2), 25–33.

Michon, P.-E., & Denis, M. (2001). When and why are visual landmarks used in giving directions? In D. R. Montello (Ed.), *Spatial information theory* (pp. 292–305). Berlin: Springer.

Miller, J., & Carlson, L. (2011). Selecting landmarks in novel environments. *Psychonomic Bulletin & Review*, *18*(1), 184 – 191. doi: 10.3758/s13423-010-0038-9

Miller, J., Carlson, L., & Hill, P. (2011). Selecting a reference object. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(4), 840–850. doi: 10.1037/a0022791

Mital, P. K., Smith, T. J., Hill, R. L., & Henderson, J. M. (2011). Clustering of gaze during dynamic scene viewing is predicted by motion. *Cognitive Computation*, *3*(1), 5–24. doi: 10.1007/s12559-010-9074-z

Mitchell, M., van Deemter, K., & Reiter, E. (2010). Natural reference to objects in a visual domain. In J. Kelleher, B. Mac Namee, & I. van der Sluis (Eds.), *Proceedings of the 6th International Natural Language Generation conference* (pp. 95–104). Stroudsburg, PA: Association for Computational Linguistics.

Mitchell, M., van Deemter, K., & Reiter, E. (2013). Generating expressions that refer to visible objects. In L. Vanderwende (Ed.), *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1174–1184). Stroudsburg, PA: Association for Computational Linguistics.

Montello, D. (2005). Navigation. In P. Shah & A. Miyake (Eds.), *The Cambridge Handbook of Visuospatial Thinking.* Cambridge, MA: Cambridge University

Press.

Mooney, A. (2004). Co-operation, violations and making sense. *Journal of Pragmatics*, *36*(5), 899–920.

Moratz, R., & Tenbrink, T. (2006). Spatial reference in linguistic human-robot interaction: Iterative, empirically supported development of a model of projective relations. *Spatial Cognition and Computation*, *6*(1), 63–107. doi: 10.1207/s15427633scc0601

Myachykov, A., Thompson, D., Scheepers, C., & Garrod, S. (2011). Visual attention and structural choice in sentence production across languages. *Language and Linguistics Compass*, *5*(2), 95–107.

Nagaraja, V., Morariu, V., & Davis, L. (2016). Modeling context between objects for referring expression understanding. In *European Conference on Computer Vision.* Berlin: Springer.

Nappa, R., & Arnold, J. E. (2014). The road to understanding is paved with the speaker's intentions: Cues to the speaker's attention and intentions affect pronoun comprehension. *Cognitive Psychology*, *70*, 58–81.

Neider, M., & Zelinsky, G. (2011). Cutting through the clutter: Searching for targets in evolving complex scenes. *Journal of Vision*, *11*(14), 7–17.

New, J., Cosmides, L., & Tooby, J. (2007). Category-specific attention for animals reflects ancestral priorities, not expertise. *Proceedings of the National Academy of Sciences*, *104*(42), 16598–16603. doi: 10.1073/pnas.0703913104

Nothegger, C., Winter, S., & Raubal, M. (2004). Selection of salient features for route directions. *Spatial Cognition and Computation*, *4*(2), 113–136. doi: 10.1207/s15427633scc0402

Onishi, K., Murphy, G., & Bock, K. (2008). Prototypicality in sentence production. *Cognitive Psychology*, *56*(2), 103–141.

Ossandón, J. P., Onat, S., & König, P. (2014). Spatial biases in viewing behavior. *Journal of Vision*, *14*(2), 20. doi: 10.1167/14.2.20

Paraboni, I., Galindo, M. R., & Iacovelli, D. (2016). Stars2: a corpus of object descriptions in a visual domain. *Language Resources and Evaluation*, 1–24.

Paraboni, I., & van Deemter, K. (2014). Reference and the facilitation of search in spatial domains. *Language, Cognition and Neuroscience*, *29*(8), 1002–1017.

Paraboni, I., van Deemter, K., & Masthoff, J. (2007). Generating referring expressions: Making referents easy to identify. *Computational Linguistics*, *33*(2), 229–254.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, *42*(1), 107–123.

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics*, *27*(1), 89–110.

Pickering, M., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*(2), 169–190.

Prat-Sala, M., & Branigan, H. P. (2000). Discourse constraints on syntactic processing in language production: A cross-linguistic study in English and Spanish. *Journal of Memory and Language*, *42*(2), 168–182.

Presson, C. C., & Montello, D. R. (1988). Points of reference in spatial cognition: Stalking the elusive landmark. *British Journal of Developmental Psychology*, *6*(4), 378–381. doi: 10.1111/j.2044-835X.1988.tb01113.x

Raubal, M., & Winter, S. (2002). Enriching wayfinding instructions with local landmarks. In M. Egenhofer & D. Mark (Eds.), *Geographic Information Science* (p. 243-259). Berlin Heidelberg: Springer. doi: 10.1007/3-540-45799-217

Reiter, E., & Dale, R. (2000). *Building natural language generation systems*. MIT Press.

Richter, K.-F., & Klippel, A. (2004). A model for context-specific route directions. In C. Freksa, M. Knauff, B. Krieg-Brückner, B. Nebel, & T. Barkowsky (Eds.), *Spatial cognition IV. Reasoning, Action, Interaction* (pp. 58–78). Berlin: Springer.

Richter, K.-F., & Winter, S. (2014). *Landmarks - GIScience for Intelligent Services*. Berlin: Springer: Springer.

Rosenholtz, R., Li, Y., & Nakano, L. (2007). Measuring visual clutter. *Journal of Vision*, *7*(2), 17–17.

Ross, R. J. (2011). *Situated dialogue systems: agency & spatial meaning in task-oriented dialogue*. Bremen: CreateSpace Independent Publishing Platform.

Roth, M., & Frank, A. (2009). A NLG-based application for walking directions. In K.-Y. Su (Ed.), *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics* (pp. 37–40). Stroudsburg, PA: Association for Computational Linguistics.

Rubio-Fernández, P. (2016). How redundant are redundant color adjectives? an efficiency-based analysis of color overspecification. *Frontiers in Psychology*, *7*(153). doi: 10.3389/fpsyg.2016.00153

Sadeghian, P., & Kantardzic, M. (2008). The new generation of automatic landmark detection systems: Challenges and guidelines. *Spatial Cognition & Computation*, *8*(3), 252–287. doi: 10.1080/13875860802039257

Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge, England: Cambridge University Press.

Siegel, A. W., & White, S. H. (1975). The development of spatial representations of large-scale environments. *Advances in Child Development and Behavior*, *10*(9), 9–55. doi: 10.1016/S0065-2407(08)60007-5

Smith, A. D., Gilchrist, I. D., Cater, K., Ikram, N., Nott, K., & Hood, B. M. (2008). Reorientation in the real world: The development of landmark use and integration in a natural environment. *Cognition*, *107*(3), 1102–1111. doi: 10.1016/j.cognition.2007.10.008

Sorrows, M. E., & Hirtle, S. C. (1999). The nature of landmarks for real and electronic spaces. In F. C & M. D M (Eds.), *Spatial Information Theory. Cognitive*

and Computational Foundations of Geographic Information Science (pp. 37–50). Berlin: Springer. doi: 3-540-66365-7

Spivey, M., Tyler, M., Eberhard, K., & Tanenhaus, M. (2001). Linguistically mediated visual search. *Psychological Science*, *12*(4), 282–286.

Stivers, T., & Enfield, N. (2010). A coding scheme for question–response sequences in conversation. *Journal of Pragmatics*, *42*(10), 2620–2626. doi: 10.1016/j.pragma.2010.04.002

Stoia, L., Shockley, D. M., Byron, D. K., & Fosler-Lussier, E. (2008). SCARE: a Situated Corpus with Annotated Referring Expressions. In N. Calzolari (Ed.), *Proceedings of the Language Resources and Evaluation Conference.* Marrakech: European Language Resources Association.

Talmy, L. (1983). *How language structures space.* Berlin: Springer.

Talmy, L. (2003). *Toward a cognitive semantics.* Cambridge, MA: MIT press.

Tanaka, J., Weiskopf, D., & Williams, P. (2001). The role of color in high-level vision. *Trends in Cognitive Sciences*, *5*(5), 211–215.

Tatler, B. W., Hayhoe, M. M., Land, M. F., & Ballard, D. H. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision*, *11*(5), 5. doi: 10.1167/11.5.5.

Tatler, B. W., & Vincent, B. T. (2009). The prominence of behavioural biases in eye guidance. *Visual Cognition*, *17*(6-7), 1029–1054. doi: 10.1080/13506280902764539

Taylor, & Tversky, B. (1992). Spatial mental models derived from survey and route descriptions. *Journal of Memory and Language*, *31*(2), 261–292.

Taylor, H., & Rapp, D. (2004). Where is the donut? factors influencing spatial reference frame use. *Cognitive Processing*, *5*(3), 175–188.

Taylor, T., Gagné, C., & Eagleson, R. (2000). Cognitive constraints in spatial reasoning: Reference frame and reference object selection. In A. Butz, A. Krüger, & P. Olivier (Eds.), *American association for artificial intelligence technical report ss-00-04* (pp. 168–172). Menlo Park, CA: Association for the Advancement of Artificial Intelligence Press.

Tenbrink, T. (2005). Identifying objects on the basis of spatial contrast: An empirical study. In C. Freksa, M. Knauff, B. Krieg-Brückner, B. Nebel, & T. Barkowsky (Eds.), *Spatial Cognition IV. Reasoning, Action, Interaction* (pp. 124–146). Berlin: Springer.

Tenbrink, T. (2007). *Space, time, and the use of language: An investigation of relationships.* Berlin: Mouton de Gruyter.

Tenbrink, T. (2011). Reference frames of space and time in language. *Journal of Pragmatics*, *43*(3), 704–722.

Theeuwes, J. (1994). Endogenous and exogenous control of visual selection. *Perception*, *23*(4), 429–440.

Theune, M., Koolen, R., & Krahmer, E. (2010). Cross-linguistic attribute selection for

reg: Comparing Dutch and English. In J. Kelleher, B. Mac Namee, & I. van der Sluis (Eds.), *Proceedings of the 6th international natural language generation conference* (pp. 191–195). Stroudsburg, PA: Association for Computational Linguistics.

Tom, A., & Denis, M. (2003). Referring to landmark or street information in route directions: What difference does it make? In W. Kuhn, M. Worboys, & S. Timpf (Eds.), *Spatial Information Theory. Foundations of Geographic Information Science* (Vol. 2825, pp. 362–374). Berlin: Springer. (doi: 10.1007/978-3-540-39923-0-24)

Tom, A., & Denis, M. (2004). Language and spatial cognition: Comparing the roles of landmarks and street names in route instructions. *Applied Cognitive Psychology*, *18*(9), 1213–1230.

Tom, A., & Tversky, B. (2012). Remembering routes: streets and landmarks. *Applied Cognitive Psychology*, *26*(2), 182–193.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136.

Tversky, B. (2011). Visualizing thought. *Topics in Cognitive Science*, *3*(3), 499–535.

Tversky, B., Kugelmass, S., & Winter, A. (1991). Cross-cultural and developmental trends in graphic productions. *Cognitive Psychology*, *23*(4), 515–557.

Tversky, B., & Lee, P. (1998). How space structures language. In C. Freksa, C. Habel, & K. F. Wender (Eds.), *Spatial Cognition* (pp. 157–175). Berlin: Springer.

Tversky, B., Lee, P., & Mainwaring, S. (1999). Why do speakers mix perspectives? *Spatial Cognition & Computation*, *1*(4), 399–412.

van Deemter, K., Gatt, A., van Gompel, R., & Krahmer, E. (2012). Toward a computational psycholinguistics of reference production. *Topics in Cognitive Science*, *4*(2), 166–183.

van Deemter, K., van der Sluis, I., & Gatt, A. (2006). Building a semantically transparent corpus for the generation of referring expressions. In N. Colineau, C. Paris, S. Wan, & R. Dale (Eds.), *Proceedings of the fourth international natural language generation conference* (pp. 130–132). Stroudsburg, PA, USA: Association for Computational Linguistics.

van Der Sluis, I., & Krahmer, E. (2004). The influence of target size and distance on the production of speech and gesture in multimodal referring expressions. In S. Kim & D. Youn (Eds.), *Proceedings of international conference on spoken language processing interspeech* (pp. 1005–1008). European Speech Communication Association: Korea.

van Gompel, R., Gatt, A., Krahmer, E., & Deemter, K. (2012). PRO: A computational model of referential overspecification. In G. Egidi, U. Hasson, R. Job, F. Vespignani, & R. Zamparelli (Eds.), *Architectures and mechanisms for language processing* (pp. 180–181). AMLAP: Italy.

Varges, S. (2005). Spatial descriptions as referring expressions in the maptask domain.

In G. Wilcock, C. Mellish, & E. Reiter (Eds.), *Proceedings of the 10th European Workshop on Natural Language Generation* (pp. 207–210). Stroudsburg, PA: Association for Computational Linguistics.

Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., & Saenko, K. (2015). Sequence to sequence - video to text. *CoRR*, *abs/1505.00487*.

Viethen, J., & Dale, R. (2008). The use of spatial relations in referring expression generation. In M. White, C. Nakatsu, & D. McDonald (Eds.), *Proceedings of the Fifth International Natural Language Generation Conference* (pp. 59–67). Stroudsburg, PA: Association for Computational Linguistics.

Viethen, J., & Dale, R. (2010). Speaker-dependent variation in content selection for referring expression generation. In I. Nitin & Z. Simon (Eds.), *Proceedings of the 8th Australasian Language Technology Workshop* (pp. 81–89). Sidney: Australasian Language Technology Association.

Viethen, J., & Dale, R. (2011). GRE3D7: A corpus of distinguishing descriptions for objects in visual scenes. In A. Belz, R. Evans, A. Gatt, & K. Striegnitz (Eds.), *Proceedings of the UCNLG and Eval: Language Generation and Evaluation Workshop* (pp. 12–22). Stroudsburg, PA: Association for Computational Linguistics.

Viethen, J., Dale, R., & Guhe, M. (2011). The impact of visual context on the content of referring expressions. In G. Claire & S. Kristina (Eds.), *Proceedings of the 13th European Workshop on Natural Language Generation* (pp. 44–52). Stroudsburg, PA: Association for Computational Linguistics.

Vogels, J., Krahmer, E., & Maes, A. (2013). When a stone tries to climb up a slope: the interplay between lexical and perceptual animacy in referential choices. *Frontiers in Psychology*, *4*. doi: 10.3389/fpsyg.2013.00154

Vorwerg, C., & Tenbrink, T. (2007). Discourse factors influencing spatial descriptions in English and German. In T. Barkowsky, M. Knauff, G. Ligozat, & D. R. Montello (Eds.), *Spatial cognition v. reasoning, action, interaction* (pp. 470–488). Springer.

Westerbeek, H., Koolen, R., & Maes, A. (2015). Stored object knowledge and the production of referring expressions: The case of color typicality. *Frontiers in Psychology*, *6*(935). doi: 10.3389/fpsyg.2015.00935

Westerbeek, H., & Maes, A. (2013). Route-external and route-internal landmarks in route descriptions: Effects of route length and map design. *Applied Cognitive Psychology*, *27*(3), 297–305.

Wither, J., Au, C., Rischpater, R., & Grzeszczuk, R. (2013). Moving beyond the map: Automated landmark based pedestrian guidance using street level panoramas. In M. Rohs & A. Schmidt (Eds.), *Proceedings of the 15th international conference on Human-Computer Interaction with Mobile Devices and Services* (pp. 203–212). New York: ACM.

Wolfe, J. M. (1994). Visual search in continuous, naturalistic stimuli. *Vision research*,

34(9), 1187–1195.

Wolfe, J. M., & Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience*, *5*(6), 495–501.

Yoon, S. O., Koh, S., & Brown-Schmidt, S. (2012). Influence of perspective and goals on reference production in conversation. *Psychonomic Bulletin & Review*, *19*(4), 699–707.